MASSACHUSETTS TECHNOLOGY COLLABORATIVE

MassTech AI Hub Data Commons Collaborative Consulting Services RFP

Request For Proposal Response

RFP #: 2026-JAII-01
Due Date: 10/07/2025

Archetype

Attachment A
Application Cover Sheet

Prepared for and Submitted to:

Massachusetts Technology Collaborative (MassTech)
75 North Drive, Westborough, MA 01581-3340
Procurement Team Leader: Sabrina Mansur
Email: proposals@masstech.org

Name of Respondent: Archetype Consulting, Inc.			
Mailing Address 180 Canal St, Suite 600	**City/Town** Boston	**State** MA	**Zip Code** 02114-1804
Telephone (802) 598-8784	**Fax**		**Web Address** **https://archetypeconsulting.com/**
Primary Contact for Clarification Josh Kreiger, Sr. Vice President - Public Sector		**Primary Contact E-mail Address** JKreiger@archetypeconsulting.com	
Authorized Signatory Andy Schlosberg, Chief Operating Officer		**Authorized Signatory E-mail Address** aschlosberg@archetypeconsulting.com	
Legal Status/Jurisdiction Massachusetts Corporation		**Respondents W-9:** **Respondents UEI No.:** LHQUPLM1X877	

USE AND DISCLOSURE OF DATA: This communication is the property of Archetype Consulting, Inc. and may contain confidential or privileged information. The information contained herein shall not be disclosed outside the Government and shall not be duplicated, used, or disclosed, in whole or in part, for any purpose other than to evaluate this proposal, provided that if a contract is awarded to this offeror as a result of, or in connection with, the submission of this data, the Government shall have the right to duplicate, use, or disclose the data to the extent provided in the contract. This restriction does not limit the Government's right to use the information contained in this document if it is obtained from another source without restriction.

Table of Contents

1.0 EXECUTIVE SUMMARY 1
2.0 PROPOSED APPROACH OF SERVICES 3
4.0. KEY PERSONNEL 44
5.0. PAST PERFORMANCE 46
6.0. CLOSING STATEMENT 52
7.0. APPENDIX A: KEY PERSONNEL RESUMES 52
APPENDIX B: RFP REQUIREMENTS MAPPED TO EPICS BASED ON PERCENT COMPLETION 62
APPENDIX C: ARCHETYPE'S W-9 65
APPENDIX D: ATTACHMENT B - AUTHORIZED RESPONDENT'S SIGNATURE AND ACCEPTANCE FORM 66

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Archetype is the Commonwealth's trusted partner leading statewide initiatives for the Data Intelligence Platform (DIP), Enterprise Data Governance Program, and Enterprise Analytics and AI Readiness Framework. Working across EOTSS, ODIA, and multiple secretariats, Archetype delivers measurable results through proven frameworks and strategic partnerships with SMX, AWS, Snowflake, data.world, and ServiceNow. Together, we will deliver the MassTech AI Hub Data Commons Collaborative (DCC) as a governed, FAIR, and scalable platform that accelerates AI development and positions the Commonwealth of Massachusetts as a national leader in responsible AI and data-driven government innovation.

1.0 Executive Summary

Archetype Consulting, Inc. (Archetype) is a Boston-based small business and long-standing partner to the Commonwealth. Archetype is a trusted leader in enterprise data modernization, governance, analytics, and AI readiness. With nearly two decades of public sector experience, Archetype supports the Commonwealth's strategic data and AI initiatives alongside the **Executive Office of Technology Services and Security (EOTSS)** and the **Office of Data Innovation and Analytics (ODIA).** Our mission is to unify governance, analytics, and AI readiness into a scalable framework that connects agencies, strengthens data access, and builds public trust in data-driven decision-making, directly supporting MassTech's vision for the AI Hub Data Commons Collaborative (DCC).

Teaming for Success

To expand delivery capability and ensure statewide scalability, Team Archetype brings together Archetype and SMX supported by strategic partners AWS, Snowflake, data.world and ServiceNow. This unified team combines governance leadership, technical delivery, and enterprise-scale platform expertise to deliver the AI Hub DCC.

The DCC architecture leverages the Commonwealth's enterprise platforms and strategic partnerships to meet MassTech's RFP priorities for security, scalability, and innovation:

- **SMX**: ***Led the first AWS deployment of a federal government compute workload***, establishing expertise in secure cloud transformation. As a certified Snowflake and AWS Premier Consulting Partner, SMX supports EOTSS through management of the Snowflake Enterprise License Agreement (ELA) for cost-effective, enterprise-wide access to the Snowflake data platform. SMX is also an approved vendor under EOTSS RFR 16-21, delivering AWS and related services across Commonwealth agencies.
- **Snowflake:** The ***Commonwealth's Enterprise Data Platform***, enabling high-performance analytics and secure data sharing across agencies. Snowflake provides elastic scaling, built-in governance controls, and AI-ready architecture for real-time analytics and cross-agency collaboration.
- **AWS:** The Commonwealth's cloud infrastructure foundation, delivering scalability, FedRAMP-aligned security, and compute power for AI workloads. AWS supports both the DCC and the ***EOTSS Secure AI Sandbox***, enabling agencies to safely test and deploy AI use cases within controlled environments.
- **ServiceNow:** The governance automation platform within the Data Intelligence Platform, embedding policy enforcement, stewardship workflows, and auditability throughout the data lifecycle. This ensures consistent compliance and transparency across secretariats.
- **data.world:** A knowledge graph-powered metadata catalog that enhances data discovery, FAIR compliance, and semantic understanding, connecting datasets and governance policies into a contextual knowledge framework for ethical AI and informed decision-making.

Together, these technologies and partnerships ensure the DCC aligns with the Commonwealth's enterprise standards, delivering a secure, explainable, and future-ready data environment that meets MassTech's RFP requirements for metadata management, synthetic data, and AI fairness tools.

Our Vision

The Massachusetts AI Hub DCC represents the next stage in the Commonwealth's data and AI evolution. Team Archetype proposed solution will provide a secure, ethical, and collaborative foundation for data sharing, analytics, and AI innovation across agencies, academia, and industry partners. Our approach focuses on three pillars:

Use or disclosure of data contained on this page is subject to the restriction on the title page.

- **Governance and Data Access:** Create a unified portal and catalog for dataset discovery, curation, and secure sharing with metadata-driven compliance and data equity.
- **AI Enablement and Transparency:** Deliver synthetic data capabilities, fairness dashboards, and explainable AI tools to ensure responsible use of advanced analytics.
- **Sustainability and Collaboration:** Implement a Commons Credits System that rewards participation and collaboration to ensure the DCC's long-term vitality.

These outcomes support MassTech's goals to promote responsible AI, expand access to high-quality data, and foster ethical, transparent innovation across the Commonwealth.

Commonwealth Leadership and Proven Delivery

Team Archetype's leadership has been instrumental in translating the Commonwealth's Data Strategy into measurable statewide results. Our ongoing work across EOTSS, ODIA, and multiple secretariats provides the operational foundation for the DCC.

- **EOTSS: Enterprise Data Governance and Intelligence Platform:** leading the Data Intelligence Platform (DIP) and Enterprise Data Governance Program, standardizing metadata, stewardship, and governance roles across agencies.
- **Statewide Secretariats and Agencies:** Team Archetype supports EOED, EOE, EOHLC, LWD, OVA, and the MBTA, providing tailored data architecture, analytics, and governance programs that advance EOTSS's federated data model, improve data literacy, and enhance decision making transparency.

Our AI Implementation for the Commonwealth of Massachusetts

Our Team is well suited to continue to support the Commonwealth on its AI journey. Archetype & SMX supported the AI Implementation Program and has provided a modular architecture and tooling framework developed with AWS and Snowflake to create a standard operating environment (**approved by EOTSS**) for rapid deployment, interoperability, and skill transfer across agencies that directly complements the AI Hub DCC.

Program Components:

- **AI Readiness Assessments**: Evaluate infrastructure, governance, and data maturity to guide agency-level AI enablement.
- **AI Use Case Development**: Identify and prioritize high-value, low-risk AI opportunities that deliver measurable improvements in efficiency and public outcomes.
- **Governance and Ethical Oversight**: Apply a Commonwealth aligned AI Governance Framework to maintain transparency, privacy, and accountability across all projects.
- **Capability Building and Workforce Enablement**: Expand AI fluency and self-sufficiency through structured training, role alignment, and managed transition models.

Frameworks and Accelerators for Impact

Archetype's delivery frameworks are purpose-built to **accelerate transformation** and deliver outcomes that are **scalable, measurable, and sustainable** in alignment with the Commonwealth's goals.

- **ICON Methodology (Identify, Calculate, Optimize, Navigate):** A structured and iterative approach connecting strategy to operations, ensuring stakeholder alignment, measurable value delivery, and long-term success.
- **SnowFlow Accelerator:** A proprietary toolset that streamlines data integration and optimization within Snowflake, reducing delivery timelines and ensuring compliance.
- **Modern Data Analytics & Reporting (MDAR) Accelerator:** A standardized framework for performance analytics and reporting that enhances transparency, visibility, and data-driven decision-making across agencies.
- **Zero Trust Architecture (ZTA)**: Built upon ***Archetype's EDM Platform Zero Trust Assessment and AI Readiness Blueprint***, this framework operationalizes CISA and NIST-aligned Zero Trust principles across the Commonwealth's data ecosystem. It establishes secure, auditable, and explainable foundations for responsible AI adoption while strengthening cybersecurity posture, compliance readiness, and data-sharing integrity.

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Our Commitment to the Commonwealth
As a Massachusetts-based firm, Archetype combines deep local expertise with measurable success in governance, analytics, and AI modernization programs across the Commonwealth. Through partnerships with EOTSS, ODIA, and multiple secretariats, Team Archetype has improved transparency, interoperability, and readiness for responsible AI. Our proven results ensure the DCC's delivery will be practical, secure, and high impact.

Team Archetype is committed to delivering a secure, ethical, and scalable DCC that embeds governance, transparency, FAIR principles, and AI fairness at every level of design. Leveraging trusted technologies, including AWS, Snowflake, ServiceNow, and data.world, Team Archetype will ensure seamless metadata management, secure synthetic data generation, and responsible AI enablement. This collaboration will continue to strengthen Massachusetts' position as a national leader in responsible AI, ethical data sharing, and data-driven government innovation, creating a connected and transparent data ecosystem that fulfills the goals of MassTech's RFP.

2.0 Proposed Approach of Services

A Feedback-Driven Approach

Team Archetype's proposed approach fully aligns with the RFP's Section 3.0: Program Phases and Key Milestones while enhancing the Commonwealth's ability to achieve value throughout the project lifecycle. We are not deviating from the requested structure; rather, we are building upon it through an iterative delivery model that increases transparency, accelerates feedback loops, and improves adoption readiness.

We will continue to follow the RFP's five defined phases:

1. Initiation & Design
2. Infrastructure & Setup
3. Development
4. Testing & Governance
5. Rollout & Adoption

However, instead of approaching these sequentially for all deliverables, we recommend applying them in three focused Epics, each representing a logical segment of functionality. This allows for faster validation, continuous stakeholder engagement, and incremental value delivery throughout execution. The three Epics include:

- **Epic 1 – Data Catalog** focuses on establishing the foundational architecture for data discovery and governance, implementing metadata management, access control, dataset publishing, and integrated search capabilities to ensure transparency and interoperability.
- **Epic 2 – Credits & Upload** introduces the operational and behavioral incentive layer, developing the centralized ledger, credit-earning and spending mechanisms, and submission workflows that encourage efficient participation and equitable data use.
- **Epic 3 – Synthetic Data** delivers privacy-preserving innovation through synthetic data generation, validation, and fairness tooling, enabling safe data experimentation and public use without compromising confidentiality.

Realized Benefits of Delivering Across the Three Epics

The phased epic delivery approach – Data Catalog, Credits & Upload, and Synthetic Data – ensures the Commonwealth achieves accelerated value realization, strong user engagement, and sustainable operational

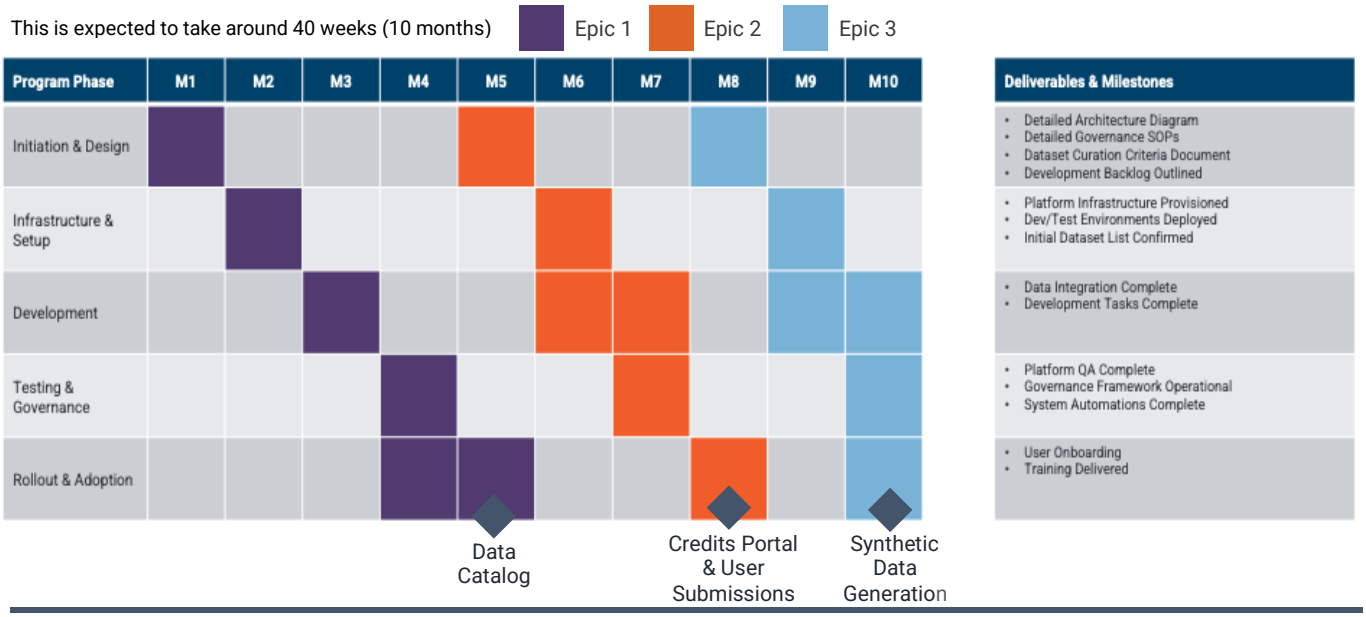
Use or disclosure of data contained on this page is subject to the restriction on the title page.

performance. Each Epic contributes distinct yet complementary outcomes that, together, create a cohesive, secure, and continuously improving Data Commons ecosystem.

1. **Accelerated Time-to-Value and Scalable Adoption**
   Deploying the Data Catalog first allows immediate access to organized datasets, searchable metadata, and governed access controls – enabling agencies to derive insights early while the platform evolves. This staged rollout delivers incremental functionality that produces value within months and allows each new feature to build on proven foundations.
2. **Functional Depth with Managed Information Flow**
   By sequencing features across multiple Epics, users receive targeted functionality and training at a digestible pace, reducing information overload and improving comprehension. This ensures adoption is deep rather than shallow – users understand and use what's available effectively before new capabilities are introduced.
3. **Behavioral Incentives and Continuous Engagement**
   The Credits & Upload Epic embeds a participation framework that motivates responsible sharing and rewards collaboration. Coupled with consistent communication on upcoming features, users see that their feedback shapes platform evolution—building trust, excitement, and a sense of shared ownership in the Commonwealth's data future.
4. **Data Privacy, Fairness, and Responsible Innovation**
   Through the Synthetic Data Epic, agencies can safely experiment with realistic but privacy-protected datasets, accelerating AI and analytics innovation without compromising confidentiality. This capability directly supports equitable access to data-driven innovation across policy domains, research, and economic development.
5. **Automation, Efficiency, and Reduced Maintenance Burden**
   Data usage patterns and site analytics captured across epics identify repetitive or manual maintenance tasks that can be automated via agents and deterministic workflows. This enhances operational efficiency, reduces human error, and enables administrators to focus on higher-value governance and strategy activities.

In summary, this three-Epic structure transforms the Commonwealth's Data Commons into an agile, user-driven, and future-proof platform, delivering faster adoption, improved functionality, measurable efficiencies, and a continuously engaged community of data users and stewards.

Figure 1. Timeline and Milestones



Use or disclosure of data contained on this page is subject to the restriction on the title page.

3.0 Technical Approach & Solution Design

Scope and Goals Confirmation

During the initial three weeks, our team will conduct a **Scope Alignment Workshop** with Commonwealth stakeholders to confirm project objectives, success measures, and dependencies across agencies. The outcome of this effort will be a jointly approved **Project Charter** that defines the mission, measurable objectives (e.g., FAIR data compliance, user adoption metrics), and governance boundaries.

This phase ensures the Commonwealth and our team maintain a unified understanding of desired outcomes—particularly around the Data Commons' core capabilities in cataloging, dataset integration, synthetic data generation, and credit-based engagement.

Technical Design and Governance Framework Development
Following charter approval, the project team will produce a **Detailed Technical Design and Governance Framework** that defines:

- **System Architecture:** Cloud infrastructure design leveraging AWS and Snowflake environments managed through EOTSS-approved environments.
- **Metadata and Catalog Standards:** Alignment with common schemas such as **Dublin Core, DCAT, and Schema.org** metadata schemas to support interoperability and FAIR compliance.
- **Security and Access Controls:** Integration of Entra ID (Azure AD) for single sign-on (SSO), role-based access control (RBAC), and data-tiered permissions consistent with Commonwealth security policies.
- **Governance Playbook:** Definition of stewardship roles, approval workflows, and lineage tracking mechanisms.

Dataset Inclusion Criteria and Validation
Archetype will lead a **Dataset Qualification and Prioritization Workshop** to develop inclusion criteria that align with the Commonwealth's open data and privacy requirements. Criteria will consider **data quality, provenance, sensitivity classification, and public value**, ensuring only datasets with validated lineage and complete metadata are onboarded into the DDC.
Once approved, the Core Team will compile a **Validated Dataset Inventory**, including hosted datasets (e.g., Massachusetts public data, MGHPCC datasets) and approved external datasets. This inventory will form the foundation for the MVP catalog and will evolve as additional datasets are integrated during subsequent build-outs.

Deliverables and Outcomes
By the end of the **Initiation & Design Phase**, the following deliverables will be completed and approved:

- **Project Charter and RACI Matrix**
- **Approved Technical Architecture and Governance Framework**
- **Dataset Inclusion Criteria and Stewardship Policy**
- **Validated Preliminary Dataset List**

This approach ensures a **strong, structured foundation** for the Commonwealth's AI Hub Data Commons Collaborative—balancing rapid mobilization with deliberate, standards-based governance to support sustainable scaling as we transition through the following four Epics.

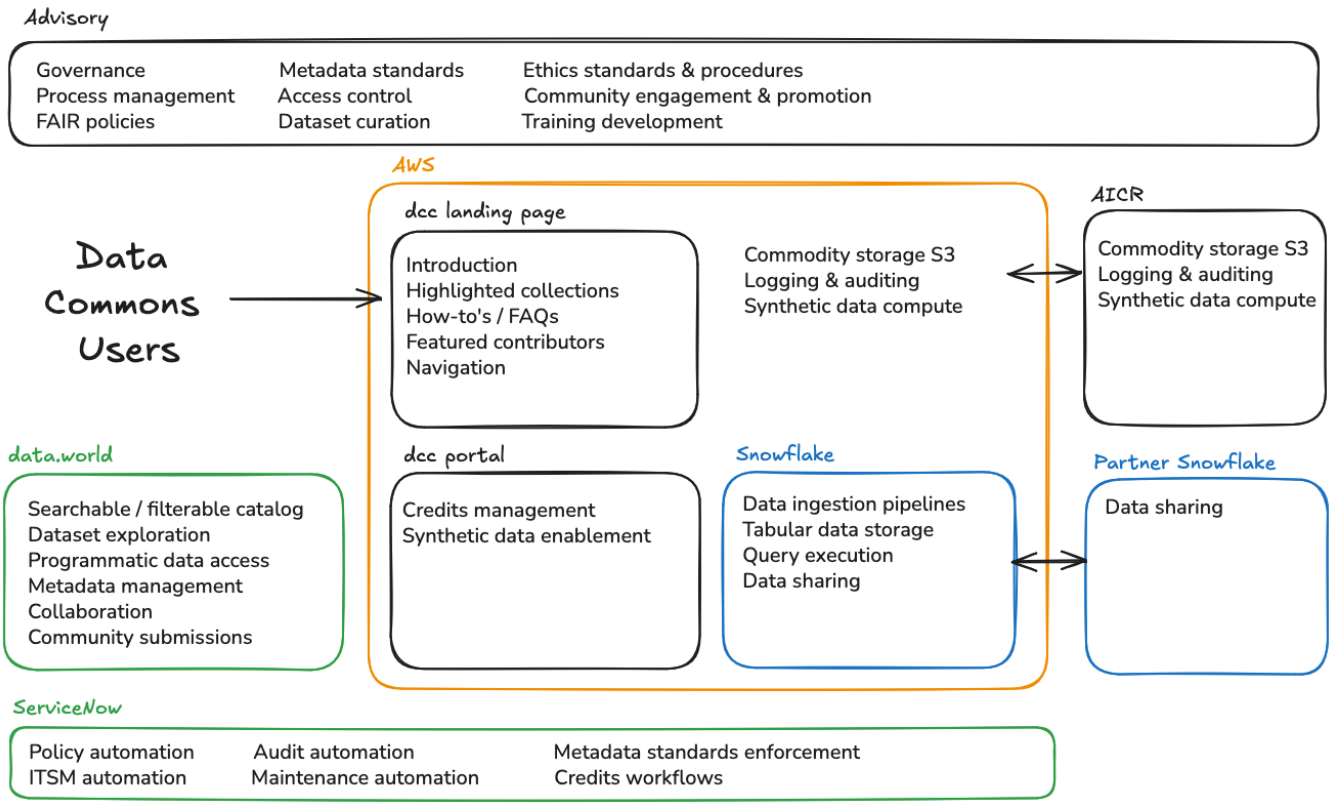
Technologies Leveraged During Implementation

The AI Hub Data Commons Collaborative (DCC) is built on a **secure, scalable, and interoperable technology ecosystem** that unifies the Commonwealth's data, governance, and AI capabilities. As shown in the **Figure 2**

Use or disclosure of data contained on this page is subject to the restriction on the title page.

below, the solution integrates **Snowflake**, **AWS**, **ServiceNow**, and **data.world** into a cohesive architecture that enables seamless data sharing, FAIR compliance, and responsible AI innovation. Snowflake serves as the Commonwealth's enterprise data platform for secure, high-performance analytics; AWS provides the scalable, FedRAMP-aligned cloud infrastructure to support compute-intensive AI workloads; ServiceNow embeds governance automation and stewardship workflows across the data lifecycle; and data.world delivers the semantic catalog and metadata intelligence to ensure discoverability, lineage, and policy alignment. Together, these technologies form the foundation for a **FAIR-compliant, AI-ready environment**, strengthened by **AI Compute Resource Infrastructure** (AICR) and Archetype's advisory services—providing change management, governance policy development, and value realization support to ensure sustainable adoption across agencies.

Figure 2. Overview of Proposed Technologies



The following technologies were carefully selected to fulfill the requirements of this project with a bias towards interoperability, leveraging of open standards, best of breed, long-term maintenance, and time-to-value.

Purpose	Technologies	Rationale	Epic
Landing Page	Payload CMS	Payload CMS offers a powerful headless CMS solution built on modern web frameworks (Next.JS, Node). It is fully open source and offers the best of both worlds with a UI for developing content (like web assets, news posts, etc.) and a code-first principle for the development aspect.	#1
Data Catalog	data.world	Many of the requirements and outcomes expressed in the RFP seem to coalesce perfectly with the value proposition that data.world offers. Specifically, it provides several mechanisms natively, including: - The ability for multiple organizations to collaborate with a standardized metadata layer in an open catalog. - Interoperability components from a SQL and SPARQL workbench, native connectivity to ML and data science platforms (like Jupyter	#1

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Purpose	Technologies	Rationale	Epic
		notebooks), a robust API, multiple SDKs, multiple file formats support, and metadata collectors from a variety of sources (including S3 and Snowflake). - A modern UI with search, filtering, discovery, and admin capabilities for column-, dataset-, and project-level entities. - The ability to include instructions for use and detailed documentation for all entities - Native support for RDF/RDFS knowledge graphs to describe and relate sources of data	
Data Platform	Snowflake	Snowflake as the underlying data platform provides several unique advantages, including: - Zero-copy secure data sharing of other Massachusetts datasets (therefore the data is always up-to-date and does not accrue additional storage costs) - Highly scalable pay-for-use compute infrastructure, including a container runtime that offers both CPU- and GPU-focused deployment options - AI features via cortex functions (complete, classify, sentiment, etc), cortex search service, and cortex analyst - Native connectivity to AWS S3 and data.world used in this proposal (among other technologies)	#1
Commodity Storage	AWS S3	Offers highly durable, scalable, and secure object storage with seamless integration to other components in this proposal including data.world, Snowflake, and other AWS services.	#1
Web Hosting	AWS Amplify	Offers simplified, automatically-scaled hosting for web applications with simpler integrations with other AWS capabilities, including CloudTrail / CloudWatch	#1
Auth / SSO	Cognito	Authentication and SSO functionality with near ubiquitous integration with identity providers and downstream applications	#1
Code Repository	Github	Most ubiquitous code repository solution with existing presence at other Massachusetts entities. Offers native CI/CD capabilities via Github Actions	#1
Logging / Monitoring	AWS CloudTrail / CloudWatch	Offers robust logging and monitoring capabilities across assets deployed in AWS	#1
AI Bias & Fairness	AIF360	• Out-of-the-box group fairness metrics such as disparate impact, statistical parity, equalized odds/opportunity, calibration by group • Pre-, in-, and post-processing algorithms (reweighing, adversarial debiasing, calibrated equalized odds, etc.) enable compare-and-choose remediation. • Works across classifiers and datasets; integrates with common Python/ML stacks (scikit-learn, pandas). • Produces reproducible evaluations you can attach to model/dataset cards—supports "measure → mitigate → re-measure" evidence for reviewers.	#1
DOI	DataCite	• DOIs resolve reliably via established resolver infrastructure, ensuring long-term, citable links to datasets and versions • Native DataCite schema captures identifiers, creators, funding, provenance and versioning which boosts findability and reuse. • Broad indexing and compatibility with Schema.org/JOSN-LD improves discovery in web and academice search • Mature APISs and client libraries for minting DOIs directly from pipelines/portals; easy to embed in CI/CD and dataset workflows.	#1
Workflow Engine	ServiceNow	• End to end governance workflow, first in class enterprise workflow engine for submission and governance • Enterprise-grade records and audit, every submission becomes a first-class record with immutable history of decisions, reviewers, attachments, and timestamps meeting audit expectations.	#1
IDPS	AWS GuardDuty and WAF	• Always-on anomaly and threat detection across CloudTrail, VPC Flow Logs, DNS—catches account compromise, data exfil, and suspicious API use around Credits/Submissions.	#1

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Purpose	Technologies	Rationale	Epic
		• Zero infrastructure to run; integrates with Security Hub and alerting for fast incident response. • Complements WAF/API Gateway throttles by spotting behavior that slips past request filters.	
Antivirus	ClamAV	• Scans community submissions in S3 to block known malware and risky active content (e.g., macro-enabled docs, embedded scripts), even when "it looks like data." • Pairs with MIME/magic-byte checks and schema sanity to enforce a data-only ingestion standard. • Lightweight to operate (Lambda/EFS or ECS) and fully auditable (engine/DB versions stored with verdicts).	#1
Static Code analysis	GitHub Code QL	• Static analysis of TypeScript/Node code to catch injection, authz flaws, unsafe deserialization, etc., before deploy. • PR-gated checks with code scanning alerts → defects fixed in development; ties into our "no separate QA team" approach. • Broad rule packs and custom queries for business-logic checks	#1
DLP/PII Detectors	Amazon Macie	• Content-aware DLP over S3 staging buckets to detect PII/PHI/secrets in uploaded datasets & docs using managed + custom identifiers. • Drives policy actions: quarantine, curator review, or conditional approval based on declared sensitivity. • Findings are structured, taggable, and feed governance analytics.	#1
EPIC 2 – New Infrastructure & Technologies			
Credits Portal	Typescript, React, Vite, Node	Using Typescript, React, Vite, and Node together creates a robust environment for building scalable and maintainable web applications. This tool stack has wide adoption and is the backbone for modern web applications. • Typescript's static typing reduces bugs and makes code refactoring more predictable, enhancing long-term reliability • React's component-based architecture simplifies the creation of modular and reusable user interfaces, improving development speed and maintainability • Vite delivers incredibly fast development cycles with instant hot module replacement, increasing productivity and ensuring modern build standards • Node enables efficient server-side and build tool capabilities, streamlining development with universal JavaScript and facilitating secure, scalable deployments	#2
Ledger	AWS QLDB	• QLDB is an append-only, cryptographically verifiable ledger (SHA-256) with a built-in journal and digest APIs. Every issuance/transfer/redemption gets an indelible, query able history and proof of integrity • Fully managed, ACID transactions, PartiQL (SQL-compatible) for reads/writesso we get audit-grade guarantees without running a blockchain. • QLDB Streams → Kinesis/EventBridge for real-time APIs and dashboards, plus journal export to S3 for batch reporting	#2
Credits API	API Gateway + Lambda	• Secure governed edge: TLS, WAF, JWT authorizers, usage plan and throttling at the gateway • Lambda scales to spikes, idempotent handlers perform ACID writes to QLDB for exact-once earn/spend	#2
Operational and config/state	DynamoDB	• low-latency, serverless, easy TTL for ephemerals, no schema migrations • Needed for credit system interm application state	#2

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Purpose	Technologies	Rationale	Epic
EPIC 3 – New Infrastructure & Technologies			
Synthetic Data Engine	synthcity	• Open Source Apache license simplifies initial integration into closed loop system • Plug-in architecture allows for extensibility into new models or data loaders • Variety of existing plugin models for generating tabular, time-series, survival, and image data • Evaluation mechanisms for correctness, fairness, and privacy https://github.com/vanderschaarlab/synthcity	#3
Synthetic Data Portal Frontend	Typescript, React, Vite, Node	Using Typescript, React, Vite, and Node together creates a robust environment for building scalable and maintainable web applications. This tool stack has wide adoption and is the backbone for modern web applications. • Typescript's static typing reduces bugs and makes code refactoring more predictable, enhancing long-term reliability • React's component-based architecture simplifies the creation of modular and reusable user interfaces, improving development speed and maintainability • Vite delivers incredibly fast development cycles with instant hot module replacement, increasing productivity and ensuring modern build standards • Node enables efficient server-side and build tool capabilities, streamlining development with universal JavaScript and facilitating secure, scalable deployments	#3
Synthetic Data Backend API	AWS API Gateway	Secure governed edge: TLS, WAF, JWT authorizers, usage plan and throttling at the gateway, interaction with HPC AICR resources including Docker API	#3

Requirements Fulfillment

This section walks through the requirements outlined in the *MA AI Hub – Data Commons Collaborative (DCC) Requirements* document to demonstrate how our solution meets or exceeds each need. The requirements below are in the same order as the document; however some needs are combined into one section when the content is directly related. For navigation purposes, there are some sections that will refer back to where the requirement was addressed. Additionally, each section will give an approximate percentage of the work that will be delivered for that requirement by epic.

Webpage (Landing Page) – Requirements 5.1

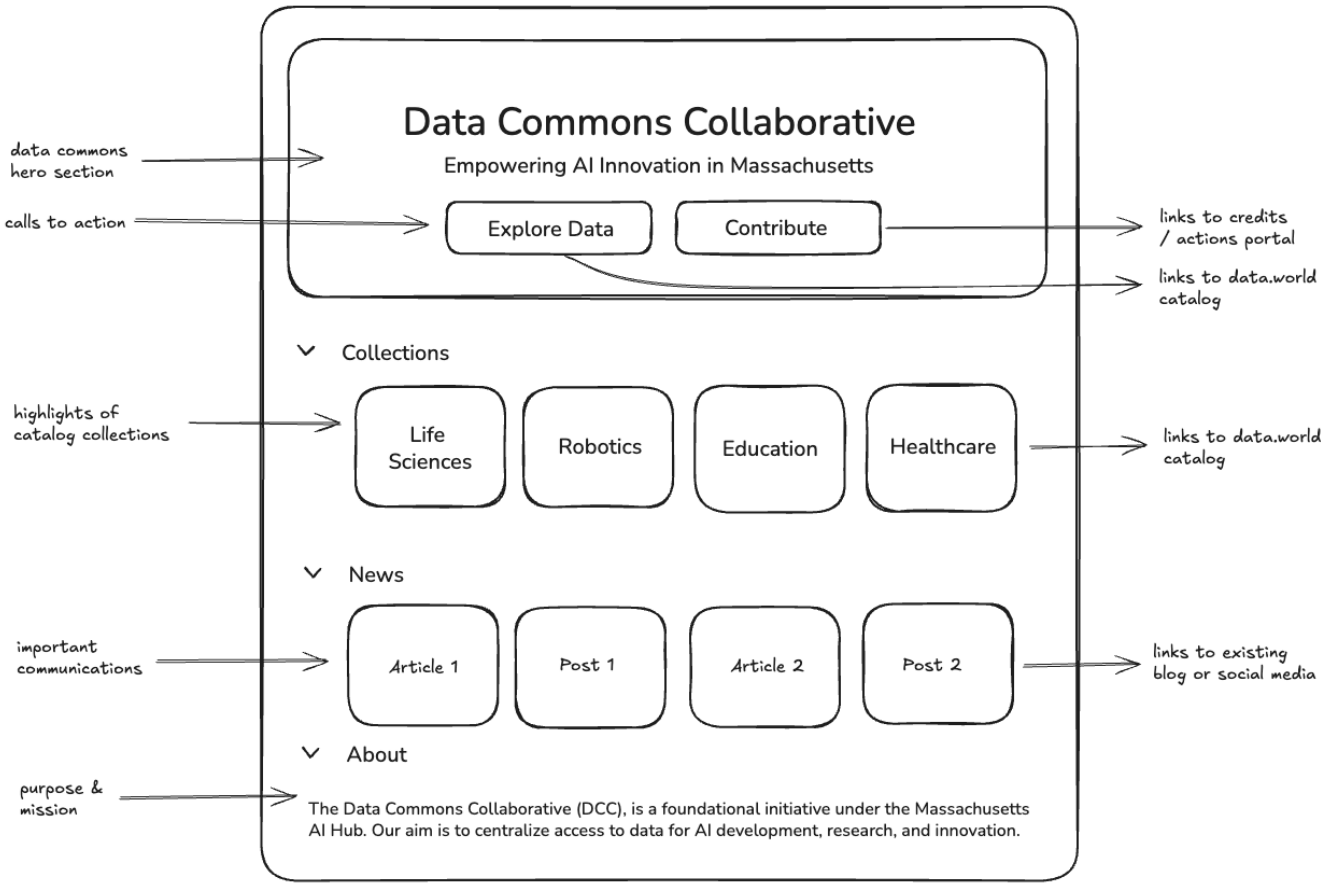
Name	#	Epic 1 (Data Catalog)	Epic 2 (Credits & Upload)	Epic 3 (Synthetic Data)
Webpage	5.1	100%		

The landing page (See Figure 3.) for the DCC will be users' first impression of the site. As such, we propose a simplified single-page application with a modern aesthetic highlighting the recommended calls to action featured in the hero banner of the page. The landing page will be managed with the open-source Payload CMS to provide the DCC team with an easy-to-use administrative back-end UI to manage the content of the landing page and communications, while providing modern development framework for our team to make custom integrations to social media and modern design packages. This will be deployed on AWS leveraging AWS Amplify, which will provide cost-effective hosting that scales up and down automatically with usage. Deploying on the AWS

Use or disclosure of data contained on this page is subject to the restriction on the title page.

environment also ensures that we can leverage CloudTrail and CloudWatch tools for logging and monitoring, as well as enable any supporting AWS resources such as DocumentDB in a secure manner.

Figure 3. Landing Page Details



In the first epic, the call-to-action section will focus on data exploration with an "Explore Data" button, but a second "Contribute" call-to-action button will feature as a part of the second epic. The "Explore Data" button will link users to the data.world catalog (described below). If the user is not yet authenticated, they will first be linked to a single sign-on page that will invite them to either log in or sign up. Newly signed-up users will have to be approved prior to gaining access to the catalog.

Scrolling down, visitors to the landing page can quickly understand the types of data available in the application with cards representing highlighted dataset collections we want to convey. This can include specific industries (Life Sciences, Robotics, Education, Healthcare) or other distinguished collections such as Massachusetts, Public, or Synthetic. These collections will link to the data.world catalog itself with the collections pre-filtered.

A dynamic section for important communications will come next which will link to content that the DCC team can create and manage through the Payload CMS backend portal. We will also feature posts from the DCC social media profiles to build up engagement with the community. This will also be a section where we can feature enablement content and educate users on new features as we roll them out.

The final section will allow for us to work together to create a narrative describing the mission of the DCC and provide context for why users should participate in leveraging the DCC for AI development, research, and

Use or disclosure of data contained on this page is subject to the restriction on the title page.

innovation. This will also be the section that we highlight the DCC's guiding principles of Transparency, Accountability, Equity, Consent and Compliance.

Dataset Directory – Requirements 5.1, 5.2

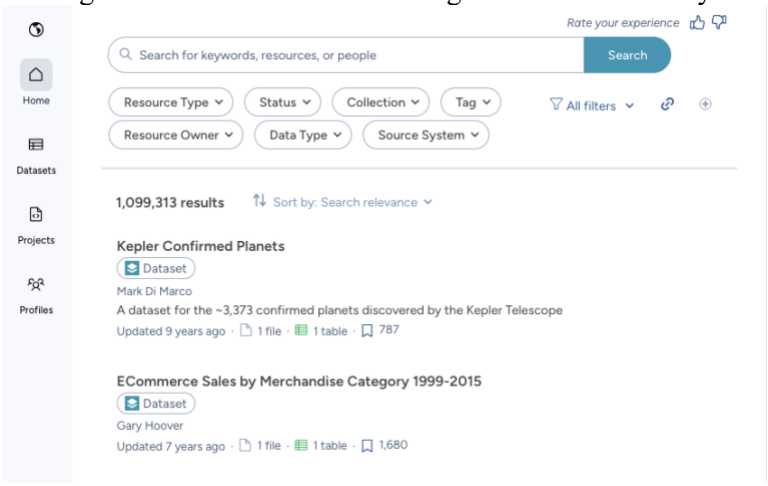
Name	#	Epic 1 (Data Catalog)	Epic 2 (Credits & Upload)	Epic 3 (Synthetic Data)
Webpage	5.1	100%		
Dataset Directory	5.2	100%		

Our team has chosen the data.world catalog to rapidly provide a solid foundation for all the functionality requested in the RFP such as metadata schema and filtering capabilities, and searchable directory with dataset cards, so our consulting team can focus on:

- curating content that will best serve the mission of the DCC and its users
- applying governance policies for quality and maintaining FAIR principles
- automations to ensure long-term maintenance is as simple and robust as possible

Below we will describe the native functionality within the data.world catalog, how it fulfills the requirements in this RFP, providing examples where applicable.

Figure 4. Data.world Data Catalog Search Functionality



The central component of the data.world catalog is the ability for users to search and filter for datasets, projects, files, or even columns (Figure 4). As you can see in the screenshot above, users can filter by collection, tag, status, resource type, resource owner, data type, or source system. These filters will be customized to match the filters outlined in the requirements, specifically by:

- Data Type
- Source
- Domain
- File Format
- License

Use or disclosure of data contained on this page is subject to the restriction on the title page.

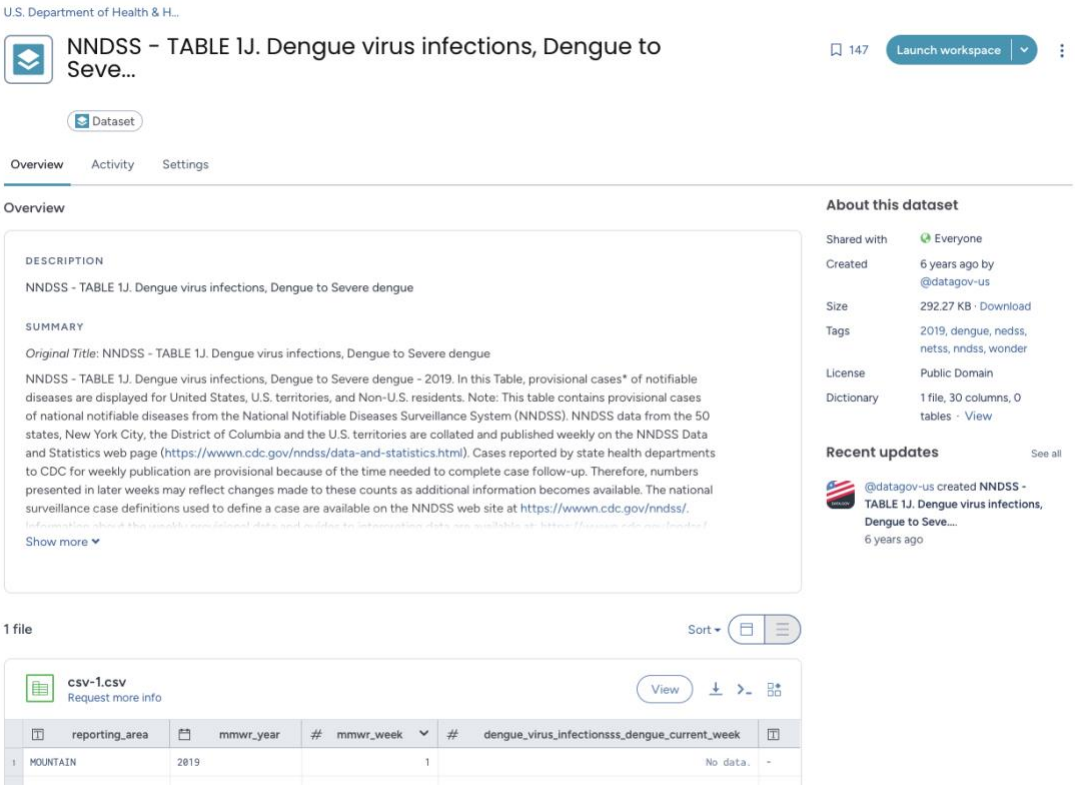
All filters are dynamically filtered based on other filter selections. Users can search quickly across all resource types and organize the results based on relevance, recency, or alphabetically. Clicking on the datasets will provide more information on the dataset including the ability to download information.

Dataset Detail Pages – Requirement 5.3

Name	#	Epic 1 (Data Catalog)	Epic 2 (Credits & Upload)	Epic 3 (Synthetic Data)
Dataset Detail Pages	5.3	100%		

The data.world catalog includes dataset detail pages (Figure 5) with metadata of the source data including the size of the data, the source, a detailed description of the information, the license, and tags that relate it to particular categories. The image below is an example of a dataset and its corresponding metadata. During the initial scoping and validation phase we will define the required metadata fields.

Figure 5. Data.world Data Catalog Dataset Detail Page Example


U.S. Department of Health & H...
NNDSS – TABLE 1J. Dengue virus infections, Dengue to Seve...
147
Launch workspace
Dataset
Overview
Activity
Settings
Overview
DESCRIPTION
NNDSS - TABLE 1J. Dengue virus infections, Dengue to Severe dengue
SUMMARY
Original Title: NNDSS - TABLE 1J. Dengue virus infections, Dengue to Severe dengue
NNDSS - TABLE 1J. Dengue virus infections, Dengue to Severe dengue - 2019. In this Table, provisional cases* of notifiable diseases are displayed for United States, U.S. territories, and Non-U.S. residents. Note: This table contains provisional cases of national notifiable diseases from the National Notifiable Diseases Surveillance System (NNDSS). NNDSS data from the 50 states, New York City, the District of Columbia and the U.S. territories are collated and published weekly on the NNDSS Data and Statistics web page (https://wwwn.cdc.gov/nndss/data-and-statistics.html). Cases reported by state health departments to CDC for weekly publication are provisional because of the time needed to complete case follow-up. Therefore, numbers presented in later weeks may reflect changes made to these counts as additional information becomes available. The national surveillance case definitions used to define a case are available on the NNDSS web site at https://wwwn.cdc.gov/nndss/.
Show more
About this dataset
Shared with
Everyone
Created
6 years ago by @datagov-us
Size
292.27 KB · Download
Tags
2019, dengue, nedss, netss, nndss, wonder
License
Public Domain
Dictionary
1 file, 30 columns, 0 tables · View
Recent updates
See all
@datagov-us created NNDSS - TABLE 1J. Dengue virus infections, Dengue to Seve....
6 years ago
1 file
Sort
csv-1.csv
Request more info
View
reporting_area
mmwr_year
mmwr_week
dengue_virus_infectionsss_dengue_current_week
MOUNTAIN
2019
1
No data.

Additionally, there is a data dictionary view that details the specific columns and their data types and descriptions. In the description (Figure 5), links to external documentation can also be provided. The metadata fields that show up in the "About this dataset" section will be customized to include a discrete field for entering external documentation links.

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Figure 6. Data.world Data Catalog Related datasets example

5 NEW YORK CITY 2019 1 No data. -
Showing 1-5 of 3,080 rows, 30 columns See all
Switch to column overview
Related datasets
NNDSS - TABLE 1J. Dengue virus in...
Dataset
@datagov-us for @us-hhs-gov
NNDSS - TABLE 1J. Dengue virus infections, Dengue to Severe dengue
Updated 6 years ago · 1 file · 1 table
NNDSS - TABLE 1J. Dengue virus in...
Dataset
@datagov-us for @us-hhs-gov
NNDSS - TABLE 1J. Dengue virus infections, Dengue to Severe dengue
Updated 5 years ago · 1 file · 1 table
NNDSS - TABLE 1J. Dengue virus in...
Dataset
@datagov-us for @us-hhs-gov
NNDSS - TABLE 1J. Dengue virus infections, Dengue to Severe dengue
Updated 5 years ago · 1 file · 1 table

Related datasets appear below the data preview (Figure 6) and are determined automatically based on the content of the description as well as lineage/provenance data.

Dataset Selection & Upload Process – Requirements 5.4, 5.6, 8.5, 9.1, 9.2

Name	#	Epic 1 (Data Catalog)	Epic 2 (Credits & Upload)	Epic 3 (Synthetic Data)
Public Data Sets	5.4	100%		
Massachusetts Public Datasets	5.6	100%		
Dataset Publishing and Versioning	8.5	100%		
Criteria for Inclusion	9.1	100%		
Suggested Initial Datasets	9.2	100%		

Hybrid Data Architecture

Before we get into the process of selecting and sourcing data, we will provide context on the mechanisms for synchronizing data between cloud resources and MGHPCC/AICR. We implement a hybrid network architecture that connects the MTC Hub AWS landing zone to MGHPCC/AICR on-premises HPC through a Transit Gateway anchored topology. High-availability AWS Site-to-Site VPN tunnels can then be established from the Transit Gateway to the MGHPCC/AICR edge to enable secure workload traffic, and we bring up AWS Direct Connect with a DX Gateway as the permanent, primary path. The VPN remains in place as failover. Routing is enforced with dedicated Transit Gateway route tables per environment, and private name resolution is delivered through Route 53 Resolver inbound/outbound endpoints and split-horizon DNS so cloud and on-prem services resolve consistently. Network segmentation isolates management, control, and data planes end-to-end.

We would standardize data movement and artifact handling. For small and medium assets, we can synchronize between S3 and MGHPCC/AICR storage via managed transfer over the private network. For large datasets, we would stage to designated storage endpoints and use parallel, resumable transfers. All staged inputs and produced

Use or disclosure of data contained on this page is subject to the restriction on the title page.

artifacts carry immutable provenance, including dataset identifiers, versions, consent/policy tags, and producing job IDs. The catalog records these linkages so downstream consumers can trace lineage and compliance.

Security controls would be enforced at multiple layers. All network links are encrypted, and we store and rotate credentials through AWS Secrets Manager with KMS-backed keys. Security groups, NACLs, and Transit Gateway route tables implement least-privilege pathways. We capture audit trails from the broker, agent, scheduler accounting, and dataset operations into centralized logging with a three-year retention policy.

Complete observability for hybrid operations can be managed through CloudWatch metrics and alarms that instrument the broker and network attachments with HPC exporter telemetry feeds into S3. Operational runbooks would then be developed that includes incident classification, SLOs for data staging, and step-by-step diagnostics for the broker/agent and connectivity layers.

Dataset Selection

For the initial epic to be successful, we need to ensure the anticipated partners (AI developers, researchers, startups, and academic institutions) use the data provided by the DCC to empower their AI innovation and research. Therefore, we envision working closely with these potential users to prioritize the sources of information that are most relevant to their research. Therefore, we recommend leveraging the following value pillars as a starting point:

- **MA One-Stop Shop** – As a Massachusetts-based organization targeting AI innovation in the Commonwealth, all the Massachusetts datasets proposed in the A.2 section of the Appendix in the RFP should be included in this initial epic. This will give a compelling value proposition for users of the site seeking any Massachusetts-related data to leverage the DCC as a one-stop shop for Commonwealth data.
- **Champion the Champions** – As we engage with potential users of the platform, we should identify willing participants as "champions" of the platform. The champions will be users that have immediate use cases that they are struggling to source data for and have a high level of engagement with outreach efforts. By targeting datasets that these champions will use, we can ensure we are providing immediate value to participants and can highlight their successes with the platform in joint-promotion materials. These success stories can then be used to find and promote future champions, creating a virtuous cycle.
- **Software-Provided Public Data Shortcuts** – Because we are leveraging AWS, Snowflake, and data.world, we can quickly integrate open-licensed datasets from these providers, which already provide descriptions, data dictionaries, and other metadata that we can quickly integrate into our platform.
- **Learn From Searches & Clicks** – In this initial phase, we will link-only datasets from all of the suggested public sources described in Appendix A.1 of the RFP that we do not fully integrate into the platform. These link-only datasets will contain descriptions and metadata for the datasets, but we will not upload any data to the DCC, instead opting to link out to the source sites. Since we can monitor user activity within the platform, we can learn what datasets most users are interested in using and create a prioritized list to incorporate in future epics.

We will actively engage with the expected end users to ensure we are providing datasets that they will find useful. Our strategy for this engagement is to host one in-person event and two virtual events where we educate potential end users about the initiative and invite the users to share their use cases and provide feedback to our approach. The events will be comprised of the following:

- Invitees:
  - AI Developers
  - Researchers
  - Startups
  - Academic Institutions
- Agenda:
  - Introduction to DCC & Plan

Use or disclosure of data contained on this page is subject to the restriction on the title page.

- Potential datasets to leverage
- Moderated discussion on use cases
- Networking

These events will allow us to identify champions of the platform and engage with them directly to ensure we can solve their use cases and use their stories to further promote the platform.

Dataset Sourcing and Metadata Capture

For this epic, we will rely on four main methods for sourcing data. To the ultimate end users, it will be a transparently seamless experience, but there are implications for the data load process.

- **Link-Only Datasets** – we will leverage AI-powered utilities to pull metadata from dataset collections (where permitted by terms of service) and create provisional dataset entries into the data.world platform. These datasets will not pull any data into our platform. After the provisional dataset is entered into the platform, administrators will be notified to review the entries before they move into an Approved status. A ServiceNow agent will be set up to validate the links monthly to ensure they are still operational.
- **Snowflake Marketplace / Data Shares** – leverage the Snowflake Marketplace and Secure Data Sharing to acquire datasets available both on the Snowflake marketplace and through sharing with other Snowflake accounts. This offers a zero-copy clone of the information that can be made available through the data.world catalog without any extract-load-transform process. For datasets that are acquired through another Snowflake account (for example another Massachusetts agency's data), we will set up an intake standard operating procedure to facilitate the acquisition process.
- **Internal Sourcing (tabular data)** – for tabular data that we source directly into our data platform for use, we establish a repeatable data pipeline within Snowflake to acquire, load, and cleanse the data in the development environment. As a part of the development lifecycle, a pull request will be made to both the Test and ultimately Production environment with peer reviews occurring at each step of the process. CI/CD tooling will ensure that the deployment of these pipelines is consistent across environments. Once in Production, the Snowflake collector from data.world will create a provisional dataset in data.world that will be reviewed by administrators until it is ultimately Approved for general use. Automated streams will be created to unload data into S3 for synchronization with MGHPCC/AICR storage via AWS Direct Connect.
- **Internal Sourcing (large collections)** – for large datasets consisting of multimodal content that we deem important to source directly into our data platform for use, we follow a similar process to the tabular data loads but ultimately load them into AWS S3 buckets instead of Snowflake tables. Similarly, we follow the same development lifecycle and CI/CD pattern, where the AWS S3 collector form data.world creates the provisional dataset. This will then be reviewed by administrators until it is ultimately Approved for general use. Automated pipelines will be created to synchronize data from S3 to MGHPCC/AICR storage via AWS Direct Connect.

In all cases, we will want to adequately document the datasets available. For a first pass we will leverage an AI utility to capture the descriptions from sources, format them using the below structure, and notify administrators to review for approval:

- Description
  - Summary
  - How to use the dataset
  - Dataset structure
  - Provenance
  - Tips & Notes
  - Full license (where required)
- Metadata Fields
  - Status
  - License

Use or disclosure of data contained on this page is subject to the restriction on the title page.

  - Source name
  - Source type
  - Data type
  - File formats
  - Tags
  - Digital Object Identifier (DOI)
  - Owner
  - Created Date
  - Last Updated Date
  - Size
  - Source description link
  - Source download link
- Data Dictionary (for tabular datasets)
  - Column name
  - Data type
  - Description
- Automatically generated metrics (for columns)
  - Sample values
  - Count of distinct values
  - Count of non-null values
  - Count of null values
  - Minimum Length (text fields)
  - Maximum Length (text fields)
  - Mean Length (text fields)
  - Mean (numerical fields)
  - Minimum (numerical fields)
  - Maximum (numerical fields)
  - Standard Deviation (numerical fields)
  - Skewness (numerical fields)
  - Kurtosis (numerical fields)

Synthetic Data Generation Capability – Requirements 5.5, 10.1-10.3

Name	#	Epic 1 (Data Catalog)	Epic 2 (Credits & Upload)	Epic 3 (Synthetic Data)
Synthetic Data Tools Section	5.5			100%
Capability Definition and Benefits	10.1			100%
Synthetic Data Generation Requirements	10.2			100%
Synthetic Data Generation Tools Samples	10.3			100%

Hybrid Architecture

As described in the hybrid architecture section above, we will implement a hybrid network architecture that connects the MTC Hub AWS landing zone to MGHPCC/AICR on-premises HPC through a Transit Gateway anchored topology. A Job Broker API in AWS (API Gateway + containerized service) will interface with an HPC

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Docker API to initiate synthetic data generation jobs on HPC resources. The broker authenticates, authorizes, and forwards job submissions over mutually authenticated TLS to a lightweight Agent service within MGHPCC/AICR. The Agent integrates with the site scheduler to submit, monitor, and cancel jobs. Users never receive direct shell access to HPC resources; all interactions flow through the broker/agent pattern with full request/response auditing.

Complete observability for hybrid operations can be managed through CloudWatch metrics and alarms that instrument the broker and network attachments and HPC exporter telemetry feeds into S3. Operational runbooks would then be developed that includes incident classification, SLOs for job submission and data staging, and step-by-step diagnostics for the broker/agent and connectivity layers.

Synthetic Data Generation Tool

For the Synthetic Data Generation capability, we will extend the credits portal with a separate page dedicated to running this functionality. Our recommendation is to begin with one synthetic data generation library for the MVP to examine and refine usage patterns before proceeding with extending the functionality to more options. For the initial implementation of this functionality, we evaluated multiple synthetic generation tools on the following qualities:

- Open-licensed – important to ensure we can natively embed the tool with better controls on usage costs
- Model configuration – should have multiple models to choose from to support a variety of use cases
- Multimodal capabilities – must be able to at least perform tabular and image synthetic generation
- Extensibility – should be able to integrate new models as technology advances
- Evaluation capabilities – should be able to monitor models for correctness, fairness, and privacy
- Community & maintenance – project should be actively maintained by community to ensure long-term viability

After evaluating multiple synthetic data generation platforms, **Synthcity emerged as the only solution that fully met—and in many areas exceeded—all evaluation criteria**, specifically:

- Uses an Apache-2.0 license
- Has many pre-built model plugins and the ability to customize more
- Has multiple configurable tabular, time-series, survival, and image generation models
- Uses a plugin-based architecture to manage generative models for extensibility and portability
- Has mechanisms for correctness, fairness, and privacy evaluation native to the platform
- Has an active community with pull requests as recent as August 2025

Synthetic Data Generation Capabilities

To enable the Synthetic Data Generation capability, we will extend the credits portal to feature a synthetic data generation page, where users can use their credits to initiate a synthetic generation job. The first time a user logs on to the page, they will be presented with the option to request access to the synthetic generation capability. After a submitting a form detailing their potential use cases, a ServiceNow workflow will trigger an administrator to approve the request for access. The administrator will be asked to review the credits the user has collected/purchased, and the use cases provided before approving the request.

Once the user has access to the synthetic data generation capability, the landing page will contain an empty table of active synthetic data generation jobs, a button to initiate a new job, and links to tutorials on how to perform the synthetic data generation jobs and how the credits system works. Initiating a new job brings up a new form to begin the synthetic data generation process where the user can input the following configurations:

- Generation Job Name
- Generation Job Description

Use or disclosure of data contained on this page is subject to the restriction on the title page.

- Dataset(s)
  - Can be a link to the data.world catalog – no further work required
  - Upload tabular (csv or parquet) or a compressed (zip or tar.gz) file containing the information that should be used as a source
    - Files uploaded are then scanned with ClamAV for malicious material
    - Data in files are processed for schema detection
    - If successfully passes the previous scans, the file is upload to AWS S3 in a directory named the same as the synthetic job id (this will be synchronized with AICR HPC storage)
- Once dataset is uploaded and schema is parsed, the user can select a model to perform the synthetic data generation from the below list (extensible if needed)
  - *General purpose*: GAN-based (AdsGAN, CTGAN, PATEGAN, DP-GAN), VAE-based (TVAE, RTVAE), Normalizing flows, Bayesian Networks (PrivBayes, BN), Random Forrest (arfpy), LLM-based (GReaT)
  - *Time Series & Time-Series Survival generators*: TimeGAN, FourierFlows, TimeVAE
  - *Static Survival Analysis*: SurvivalGAN, SurVAE.
  - *Privacy-focused*: DECAF, DP-GAN, AdsGAN, PATEGAN, PrivBayes.
  - *Domain adaptation*: RadialGAN.
  - *Images*: Image ConditionalGAN, Image AdsGAN.
- Depending on the model selected, the user will then be able to add certain parameters including:
  - Sensitive Features (multiselect of columns from source schema)
  - Important Features (multiselect of columns from source schema)
  - Target Column (single select of columns from source schema)
  - Fairness Column (single select of columns from source schema)
  - Domain Column (single select of columns from source schema)
  - Time to Event Column (single select of columns from source schema)
  - Random State (integer)
  - Height (integer) – for image sources
  - Width (integer) – for image sources
  - Train Size (float) – for image sources
  - Generation Count (integer)

Once the job is submitted, a task will be sent to Job Broker API to begin the generation process by calling the HPC Docker API. Based on the entries provided during the submission process, environment variables will be set to guide the code in the docker container to conduct the generation procedure. This will result in a second dataset being saved in the S3 directory. After the generation procedure is complete, a separate task will kick off to perform evaluation on various metrics including

- Sanity checks (e.g. close_values_probability, distant_values_probability)
- Statistical tests (e.g. inverse_kl_divergence, chi_squared_test)
- Synthetic data quality (e.g. performance.xgb, detection_xgb)
- Privacy metrics (e.g. k_anonymization, identifiability_score)

Once the generation and evaluation tasks are completed, the status in the UI will be updated to “Ready for Submission” and the initiator will get a notification that the job is complete. A ServiceNow automation will be created to gather statistics about the task runtime (including time spent, resources used, and a usage score based on other tasks running at the same time) and apply credits usage to the initiator’s balance.

Once the user logs back into the synthetic generation job page, they will have the opportunity to download or delete the resulting dataset(s). After reviewing the outputs, they can then submit the data to the DCC for inclusion

Use or disclosure of data contained on this page is subject to the restriction on the title page.

in the data catalog. This triggers a ServiceNow automation that requests the approval of a DCC administrator to review the request. Once submitted, the data will be loaded into Snowflake or S3 and follow the standard sourcing process described above. As a part of the submission to the catalog, the data will be clearly labeled as synthetic with provenance, model parameters, and evaluation metrics included in the description.

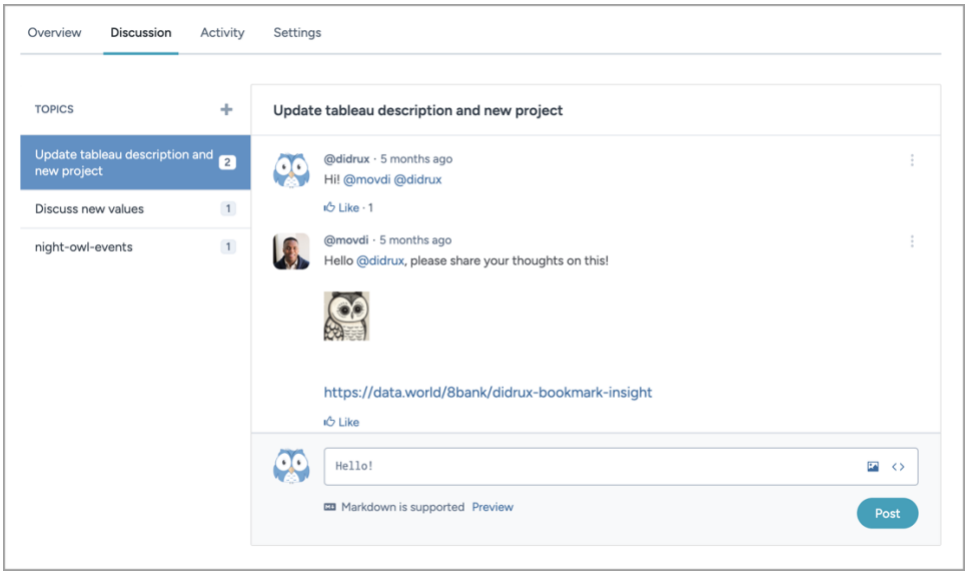
Dataset Selection Process – Requirements 5.6
Already discussed in *Dataset Selection & Upload Process – Requirements 5.4, 5.6, 8.5, 9.1, 9.2* section above. This section is here to aid in navigation of this proposal.

Community Dataset Submissions – Requirements 5.7, 8.8

Name	#	Epic 1 (Data Catalog)	Epic 2 (Credits & Upload)	Epic 3 (Synthetic Data)
Community Dataset Submissions	5.7	30% (feedback)	70% (submissions)	
Collaboration and Community Features	8.8	70%	30%	

In the first epic, we will rely on native functionality in the data.world platform to meet the requested requirements, which we detail below.

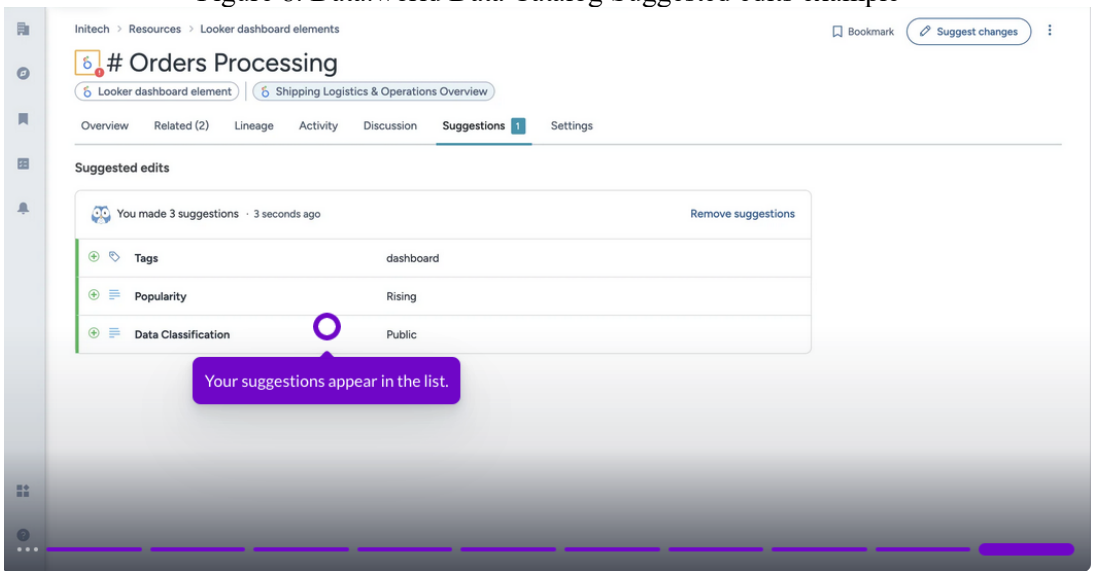
Figure 7. Data.world Data Catalog discussion topic example



The data.world catalog features collaboration as a first-class citizen. Every project and dataset has the ability for collaborators to create discussion topics (Figure 7) and provide suggestions or ask questions related to the specific entity.

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Figure 8. Data.world Data Catalog Suggested edits example



Users can also collaborate by making suggestions through the "Suggest Changes" workflow. Here instead of just creating a discussion topic to recommend a change, other participants can make the actual change and then have them approved by the project or dataset owner, see Figure 8.

Community dataset submissions will be implemented in Epic 2. These submissions will function through a guided contributor form (ServiceNow Catalog item and/or a custom portal "Contribute" page) that captures required descriptive, technical, and provenance metadata; validates links and files; and opens a curation ticket with SLAs. A provisional dataset is created in data.world for curators to enrich. On approval, the catalog entry is finalized with APIs, previews, RBAC, versioning and where applicable Digital Object Identifier readiness; the portal re-indexes automatically.

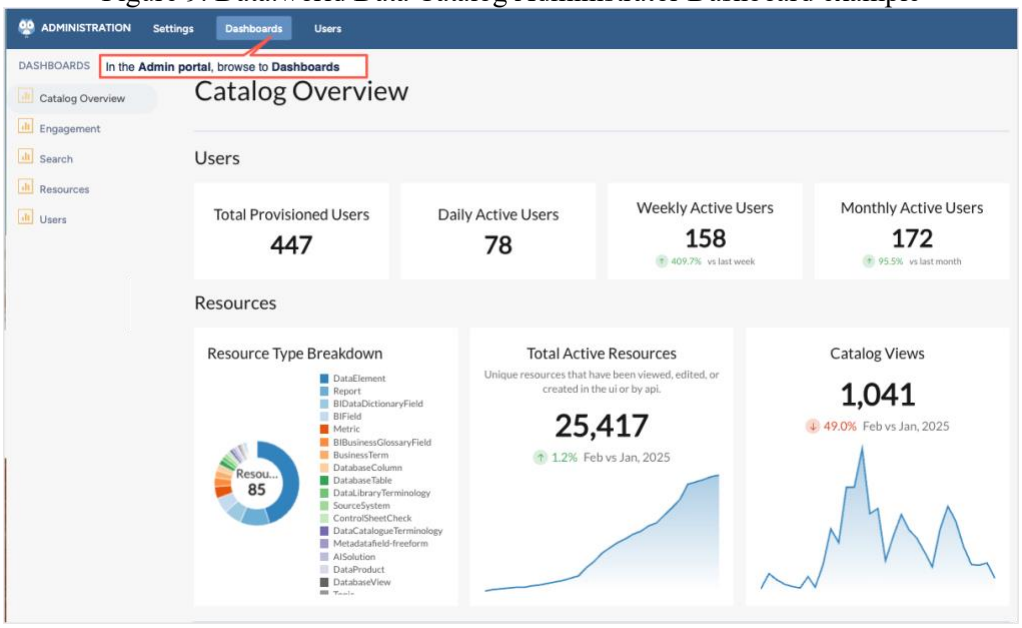
Admin Dashboard – Requirements 5.8, 8.7

Name	#	Epic 1 (Data Catalog)	Epic 2 (Credits & Upload)	Epic 3 (Synthetic Data)
Admin Dashboard	5.8	50% (dataset mgmt)	50% (submission mgmt)	
Data Usage and Metrics	8.7	100%		

There is not a specific admin portal for managing content within the data.world catalog itself, as all actions such as adding/editing/removing datasets are managed by role-based access controls assigned to the datasets themselves. Users granted with "Manage"-tier roles can add/edit/delete datasets and projects, accept proposed submissions or suggested changes, and manage role-based permissions for other users. Furthermore, admins can verify that submissions are cited properly during the approval process.

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Figure 9. Data.world Data Catalog Administrator Dashboard example



However, there is an admin portal that contains access to general user management, site-wide settings, and governance dashboards. Admins can both view and download usage data from these governance dashboards, see Figure 9 above for an example.

As described in the previous section *Community Dataset Submissions – Requirements 5.7, 8.8*, submissions from users in Epic 2 will trigger an approval workflow. Approvals run in ServiceNow with role-based routing (Curator -> Admin - > Compliance as needed), auditable decisions, and an admin dashboard for queues, link health, and role management. Publishing updates data.world and the portal index; credits are awarded upon publish.

Accessibility, Performance, Security, Responsiveness, Extensibility – Requirements 5.9, 13.2

Name	#	Epic 1 (Data Catalog)	Epic 2 (Credits & Upload)	Epic 3 (Synthetic Data)
Additional Requirements	5.9	50%	25%	25%
Data Protection	13.2	60%	20%	20%

Accessibility

Data.world incorporates several accessibility features aimed at enhancing usability for individuals with disabilities, particularly those who rely on assistive technologies. The platform supports system-level screen readers like VoiceOver (Mac) and Narrator (Windows), and it includes keyboard navigation throughout the interface. Detailed guidance is provided for keyboard-only interaction, including tab navigation, dropdown selection, and content editing via shortcuts. Data.world incorporates the standards of WCAG 2.1 Level A and

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Level AA into software development best practices and conduct periodic reviews of our software against the standard and actively work to remediate any gaps alongside our other product commits.

Performance

While data.world is engineered for responsive performance and optimized user experience, particularly for enterprise-scale deployments. Search capabilities are powered by efficient APIs and refined through ranking mechanisms that prioritize relevance, metadata quality, and engagement signals (e.g., bookmarks). Our consulting team will work with data.world support in benchmarking performance and making optimizations in our configuration to ensure the performance metrics outlined in the solicitation requirements.

Security

All interactions between all systems in this proposal are encrypted with TLS 1.2+ encryption while in transit. All information stored in data.world, AWS S3, and Snowflake are encrypted at rest, leveraging the keys provided these platform providers. If needed, customer-managed keys can be leveraged for further control of encryption standards. By default, the rate limit for API calls is initially set to 5 requests, followed by a subsequent rate of 1 request per second after a 5-second pause.

Responsiveness

The data.world catalog site implements responsive design for all the form factors mentioned in the solicitation (mobile, table, and desktop). Any custom web development will also ensure responsive design principles are incorporated into the final output.

Extensibility

Because the open community is core to the platform's identity, data.world is architected with open standards and extensibility in mind. This includes:

- **Semantic web / linked-data foundations**: RDF / OWL / SPARQL underpin the knowledge graph and metadata layer.
- **Interoperable schemas**: Support for widely adopted vocabularies (DCAT, Schema.org, PROV) to facilitate cross-catalog alignment and federated search.
- **Open APIs & SDKs**: Every dataset and metadata record is exposed via queryable APIs, enabling independent tools, visualizations, and downstream integrations.
- **Community extensibility**: Users can contribute scripts, notebooks, visualizations, and metadata improvements directly on published datasets – these enhancements feed back into the catalog, improving quality over time.

Because these technologies are "open by design," data.world does not impose proprietary silos or opaque formats. That is a critical differentiator when an organization wants a catalog that can interoperate, evolve, and invite external participation without lock-in.

Data Hosting and Access – Requirements 6.1

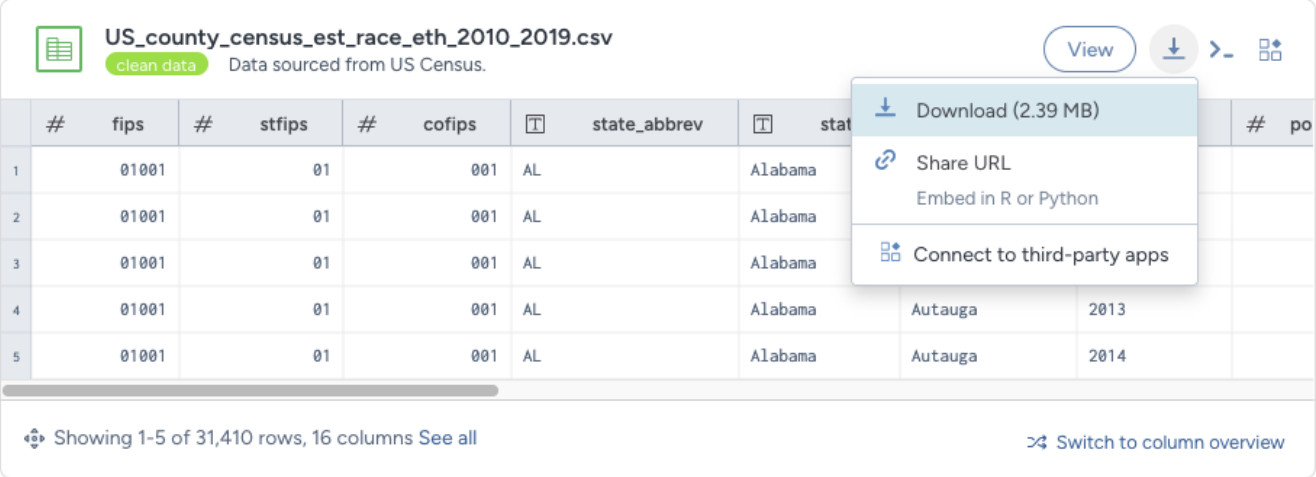
Name	#	Epic 1 (Data Catalog)	Epic 2 (Credits & Upload)	Epic 3 (Synthetic Data)
Data Hosting and Access	6.1	100%		

There are no inherent restrictions on the file types that can be uploaded to data.world catalog, but we will be enforcing specific sourcing patterns to ensure the standards we agree on are met. In addition to the sourcing pattern we define, we will also use SHACL shape files to define specific constraints for RDF metadata to meet.

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Violations of these standards will result in notifications that can be tracked down and addressed in a systematic way.

Figure 10. Data.world Data Catalog download example

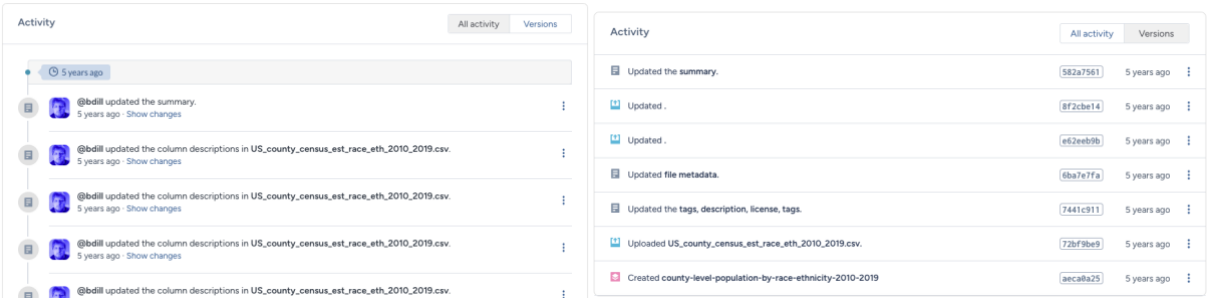


All datasets sourced into the data.world catalog are downloadable (Figure 10) into machine-readable formats or used directly via the REST API or one of the following pre-built SDKs:

- Python SDK
- R SDK
- Go SDK
- JDBC Driver

The datasets and their metadata are searchable via these tools and via the SPARQL protocol.

Figure 11. Data.world Data Catalog Activity Log example



A full activity log and version history are available with every dataset in the Activity tab of the dataset. This gives users to ensure full transparency into the history of the datasets and allows them to work off a particular version for their particular research tasks.

User Experience Enhancements – Requirements 6.2

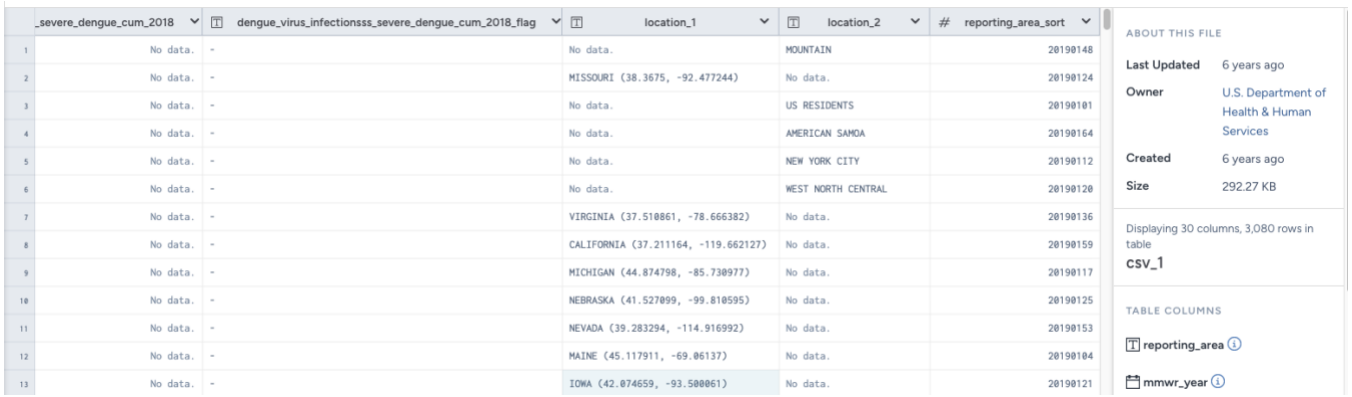
Use or disclosure of data contained on this page is subject to the restriction on the title page.

Name	#	Epic 1 (Data Catalog)	Epic 2 (Credits & Upload)	Epic 3 (Synthetic Data)
User Experience Enhancements	6.2	100%		

Although this section is labeled as "Enhancements" in the RFP, these features are already native to the data.world catalog system. Below we will go through how these requirements are implemented.

As described in the *Dataset Directory – Requirements 5.1, 5.2* section above, the data.world catalog already comes with a dataset explorer that allows users to search or filter for datasets with a variety of filters.

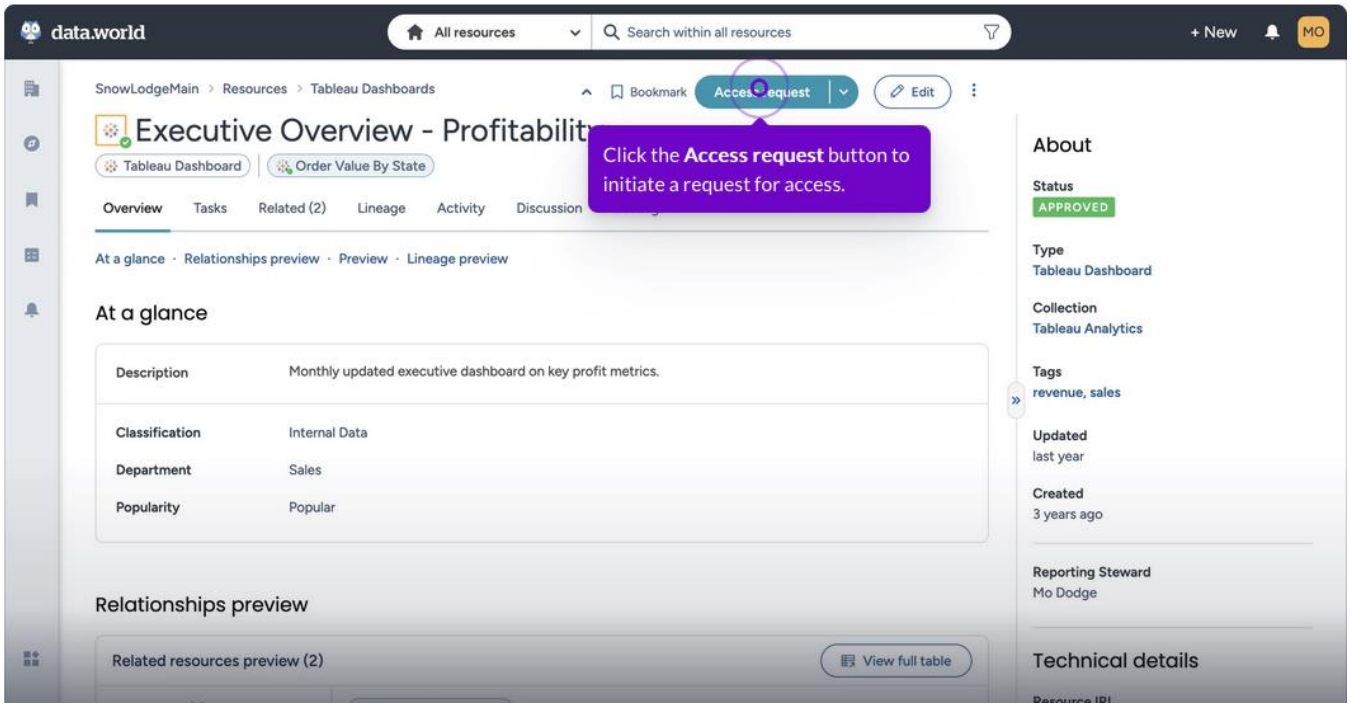
Figure 12. Data.world Data Catalog Preview functionality example


_severe_dengue_cum_2018
dengue_virus_infectionsss_severe_dengue_cum_2018_flag
location_1
location_2
reporting_area_sort
1 No data. - No data. MOUNTAIN 20190148
2 No data. - MISSOURI (38.3675, -92.477244) No data. 20190124
3 No data. - No data. US RESIDENTS 20190101
4 No data. - No data. AMERICAN SAMOA 20190164
5 No data. - No data. NEW YORK CITY 20190112
6 No data. - No data. WEST NORTH CENTRAL 20190120
7 No data. - VIRGINIA (37.510861, -78.666382) No data. 20190136
8 No data. - CALIFORNIA (37.211164, -119.662127) No data. 20190159
9 No data. - MICHIGAN (44.874798, -85.730977) No data. 20190117
10 No data. - NEBRASKA (41.527099, -99.810595) No data. 20190125
11 No data. - NEVADA (39.283294, -114.916992) No data. 20190153
12 No data. - MAINE (45.117911, -69.06137) No data. 20190104
13 No data. - IOWA (42.074659, -93.500061) No data. 20190121
ABOUT THIS FILE
Last Updated 6 years ago
Owner U.S. Department of Health & Human Services
Created 6 years ago
Size 292.27 KB
Displaying 30 columns, 3,080 rows in table
csv_1
TABLE COLUMNS
reporting_area
mmwr_year

The Data.world catalog comes pre-built with preview functionality that allows users to view a sampling of data within the application, see Figure 12 for an example. Users can also query this data directly, with pushdown queries in our selected data platform Snowflake. Pre-built queries can also be customized for each data source as starting points for users.

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Figure 13. Data.world Data Catalog Access Request example



Users can request access to resources they do not currently have access to via the Access Request workflow, see Figure 13 as an example. This creates an email and a ServiceNow ticket to have an administrator approve this access request, which triggers the required permissions flows to activate.

Common Credits System – Requirements 7.1 – 7.7

Name	#	Epic 1 (Data Catalog)	Epic 2 (Credits & Upload)	Epic 3 (Synthetic Data)
Centralized Ledger and Tracking Infrastructure	7.1		70%	30% (synth usage)
Earning Commons Credits	7.2		100%	
Spending Commons Credits	7.3		90%	10% (synth usage)
Incentivizing Efficient Usage	7.4		50%	50%
Governance and Transparency	7.5		100%	
User Experience & Access Control	7.6		100%	

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Name	#	Epic 1 (Data Catalog)	Epic 2 (Credits & Upload)	Epic 3 (Synthetic Data)
Technical Expectations	7.7		70%	30%

We will implement a centralized, auditable Commons Credits system backed by an immutable ledger (AWS QLDB) and exposed through a secure API layer using AWS API Gateway and Lambda. Every earn or spend event is recorded as an append-only transaction with a canonical schema that includes actor, organization, purpose code, dataset/model reference, policy version, timestamps, and cryptographic proofs of integrity. Transactions are ACID and idempotent, guaranteeing exactly-once posting for issuance, debits, reversals, and adjustments. Real-time read APIs and scheduled exports provide auditors and administrators with complete visibility into balances, limits, and transaction histories. The program operates as a closed loop at launch: credits are issued by policy and spent on defined resources; there is no fiat conversion or peer-to-peer trading. Governance controls enforce per-user and per-organization limits, approval workflows, and retention policies, with a full audit trail available for oversight.

We will deliver a clear, accessible user experience that makes credits understandable and actionable. The portal presents a persistent balance widget, a transaction history with filters and export, and contextual "how you earn" and "what this costs" panels embedded at the point of action (example: dataset download, premium API call, MGHPCC/AICR job submission). Before any spend, the UI displays a deterministic estimate, checks available balance, and shows the policy that applies; on confirmation, the debit is posted and reflected immediately. Role-based views support researchers, submitters/curators, and admins: researchers see balances, usage, and caps; curators see issuance tied to accepted contributions; admins manage policies, limits, and organization-level reporting.

We award Commons Credits for verifiable contributions that increase the value, quality, or responsible use of the Hub. Credits are issued automatically by a policy engine when specific events complete in the submission and curation workflow (Service Now), and every issuance posts as an immutable transaction to the ledger.

Possible earn events at launch:

1. Accepted dataset submission with required metadata and license
2. Approved metadata enrichment to an existing record
3. Validated documentation or tutorial that enables reproducibility or safe use;
4. Curator review tasks completed by delegated reviewers
5. Synthetic data package published with provenance and risk profile. Each event fires from the workflow system after approval and publishes a signed issuance to the Credits API.

We enforce rigor and anti-gaming controls. Submissions must pass automated checks and human review where required. The policy engine verifies uniqueness (no double credits for issuance of same artifact), authorship whether submitter or organization, and completeness. Issuance amounts are deterministic and published they vary by contribution type and quality tier. Credits are non-transferable, subject to per-user and per-organization caps, and ties to the contributing identity and artifact IDs for audit.

We make earning transparent in the product. The portal surfaces a persistent balance and a history tab showing each issuance with its policy, artifact link, reviewer, and timestamp. At the point of action (e.g., after a dataset is approved), users see the exact number of credits granted and the reason code. Organizations receive roll-up views so team leads can track contribution volume and earned credits against program goals.

Use or disclosure of data contained on this page is subject to the restriction on the title page.

We administer governance centrally. Program admins manage policy versions, caps, and quality tiers; they can grant or reverse credits through an auditable admin action if for example plagiarism or gaming is detected. Reversals post as compensating transactions referencing the original issuance. This model ensures credits reflect real, high-quality contributions while remaining fair, predictable, and fully auditable.

We enable credits to be spent on clearly defined, high-value actions inside the Hub, with deterministic pricing and pre-authorization. At launch for epic 2, spendable actions include:

1. access to premium or rate-limited datasets and models
2. submission of compute jobs
3. prioritized or high-throughput service tiers for portal APIs

Each action publishes a price table maintained by program admins, so the portal can display an up-front estimate and run a pre-spend check against the user's and organization's balances. If sufficient credits exist and policy constraints are satisfied, the Credits API places a provisional hold and the transaction proceeds. In Epic 3, as described in section *Synthetic Data Generation Capability – Requirements 5.5, 10.1-10.3*, credits can be further used to leverage synthetic data generation compute.

We post final debits to the ledger using exactly-once, idempotent writes when the action begins (for fixed-price items) or on completion (for metered usage). If a job fails before consuming resources, the hold is released; if it partially runs, the debit reflects only the metered portion. Every spend transaction includes actor, organization, resource identifier, policy version, timestamps, and a purpose code to support audit and reporting. Reversals such as refunds for data quality defects or policy exceptions—post as compensating transactions that reference the original debit.

Spending is governed by role- and tier-based controls and program-level limits. Users and organizations have configurable hard caps and per-action ceilings (e.g., maximum concurrent compute jobs, maximum premium API calls) enforced at the API Gateway and in the policy engine. The platform applies throttling and queueing where appropriate to protect shared services. Admins can grant temporary exceptions with an auditable override flow; such exceptions are time-boxed and tied to a specific action or project. For organizational transparency, admins and team leads see live credit posture at the org level and can set internal budgets by project or cost center.

The user experience makes spending intentional and understandable. Before a spend, the portal shows the estimated cost, the factors driving it, and the applicable policy. On confirmation, the user sees an immediate balance update and can track the transaction's status from "held" to "posted," along with links to the underlying dataset/API call or compute job. For metered items, the UI presents interim usage and the evolving estimate; on completion, the final debit, receipt, and artifacts are available for export. This design ensures credits are spent on defined resources with up-front transparency, strong enforcement, and a complete, auditable record.

We design the Commons Credits system to actively steer users toward efficient, sustainable behavior through deterministic pricing rules and automated enforcement. The policy engine applies weights and multipliers to every spend based on measurable factors: resource class, time window, queue/partition utilization, job efficiency signals, and data-locality. Off-peak submissions receive reduced credit rates, while peak-window usage and high-demand GPU types accrue a surcharge (relevant to MGHPCC jobs not Snowflake as Snowflake has fast auto scale per request capabilities). Jobs that run in designated green windows at MGHPCC/AICR aligned to underutilized capacity and energy-efficiency targets receive additional discounts.

We reward data-proximity and reuse. Training and inference executed where the dataset already lives receive a locality discount; large cross-domain transfers incur a transfer tariff in credits. When users leverage cached intermediates, shared feature sets, or approved model checkpoints, the debit reflects a lower rate than full

Use or disclosure of data contained on this page is subject to the restriction on the title page.

retraining. We maintain a shared registry of artifacts eligible for reuse and issue bonus credits for publishing reproducible pipelines that others adopt, reinforcing a virtuous cycle of efficiency.

We enforce and make these incentives transparent through the ledger and UI. Every quote and final debit lists the policy version and the exact factors that adjusted the price. Organization and program dashboards surface efficiency KPIs average utilization, off-peak adoption rate, cache-reuse rate, and avoided transfer volume so leaders can govern behavior with evidence. Administrators manage weights and windows centrally and publish scheduled changes in advance; changes take effect per policy version and are recorded immutably in the ledger for audit. This design ensures credits consistently favor efficient, carbon-aware, and cost-effective use of the platform.

We operate the Commons Credits program under a documented governance model with clear roles, controls, and auditability. A Program Owner (MassTech) sets policy and delegates day-to-day operation to a Credits Admin team with segregation of duties across policy management, issuance/grant approval, and reconciliation. All credit events (issuance, spend, adjustment, reversal) are written to an immutable ledger with actor, organization, purpose code, artifacts, timestamps, and policy version. Policies, limits, and price tables are versioned; effective dates and deprecation windows are recorded so any transaction can be re-evaluated against the exact rules in force at the time.

We enforce preventative controls before value moves. The portal and API perform pre-authorization checks (identity, role, organization affiliation), balance and limit checks (per-user and per-org daily/monthly caps), and policy validations (eligibility, completeness, uniqueness). Administrative actions that create or adjust credits (example: grants, reversals) require dual control: a maker–checker workflow in ServiceNow with electronic signatures, reason codes, and attachments. High-risk changes such as altering multipliers or introducing new spend categories require change control with stakeholder review, scheduled effective dates, and broadcast notices to users.

We maintain comprehensive monitoring, reporting, and reconciliation. Operations publishes a governance dashboard with live posture (balances, burn rate, cap utilization), exception queues (failed validations, fraud flags), and KPIs (off-peak adoption, reuse rate, reversal rate, average time to approve). The ledger streams (See QLDB streams for near real-time analytics) to analytics for daily reconciliations (issuances = debits + balance deltas) and variance analysis by organization and resource class. Quarterly, the Program Owner conducts an attestation of policy adherence and access reviews (role recertification, least-privilege checks). All logs and records are retained for at least three years and are exportable for oversight.

We provide transparent issue and dispute handling. Users can appeal debits or denials through a tracked case; investigators review the ledger entries, artifacts, and policy version, then either uphold the action or post a compensating transaction (reversal or adjustment) with rationale. Confirmed abuse (e.g., duplicate submissions, resource hoarding) triggers policy-defined sanctions from temporary caps to account suspension and is reported to program leadership. Aggregate learnings feed into policy tuning (weights, caps, qualifying events), with changes communicated in advance and applied under a new policy version.

We safeguard security and privacy throughout. Identity is federated; roles map to fine-grained API scopes; all administrative endpoints require MFA. Secrets and keys are centrally managed; all data in transit and at rest is encrypted. The governance process integrates with the broader platform's risk, legal, and privacy reviews, ensuring that credits are issued only for contributions and usage that meet licensing, consent, and ethical standards. This end-to-end model delivers a predictable, fair, auditable credits program that aligns incentives while protecting the integrity of the Hub.

Metadata Capabilities and Management – Requirements 8.1, 8.2

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Name	#	Epic 1 (Data Catalog)	Epic 2 (Credits & Upload)	Epic 3 (Synthetic Data)
Capability Definition and Benefits	8.1	100%		
Metadata Management	8.2	100%		

We adopt a "FAIR-by-default" metadata profile that blends descriptive, technical and provenance fields aligned to common schemas (Dublin Core, Schema.org, DCAT) so records are discoverable, interoperable, and consistently documented across hosted and linked assets. In practice, the submission form and curator tools enforce these fields (title, abstract, keywords, license; file formats/schemas/dates; source and processing lineage), and the catalog exposes them via APIs for programmatic reuse. This directly implements Descriptive metadata, technical metadata, provenance, and explicit support for metadata standards.

For hosted data, every dataset is published with machine-readable formats (CSV/JSON/Parquet), version identifiers, and dataset-level lineage so consumers can retrieve, integrate and cite consistently; the same metadata also powers search and browse in the public portal. We surface this profile in data.world (system of record) and reflect it in the website's index so users get fast, faceted discovery and consistent detail pages.

To make datasets citable, we include a standard citation block on each dataset page (creator, title, year, version, publisher/host, URL) and support persistent identifiers (eg. DOIs via DataCite) for stable, hosted releases. The catalog tracks usage and citation linkages over time, satisfying the RFP's emphasis on citation tracking and DOI readiness. Practically upon approval the workflow can mint or record a DOI and write it into the data sets metadata, while the catalog maintains download/access metrics and citations.

Because access controls are policy as much as plumbing, the catalog record also stores the dataset's access tier (public/restricted/internal) and any license usage constraints so the portal and APIs can enforce the right behavior at read time. With SSO/MFA and RBAC in place, restricted datasets only, restricted datasets only become accessible after an approval step; all reads and downloads are logged for audit. This represents a tight integration and coupling of both enforcement and metadata processes.

In Epic 2, we wire FAIR into the Credits & Contributions flow so that submissions are Findable, Accessible, Interoperable, and Reusable before credits are issued. Every contribution must include rich metadata (title, abstract, keywords, domains), persistent identifiers (e.g., DOI/handle), declared access tier and machine-readable license, approved open formats (CSV/JSON/Parquet/Avro) with schema/encoding, provenance and versioning, and usage notes. These checks are enforced in the ServiceNow workflow as gates; the Credits Policy Engine awards more credits for higher FAIR quality (complete schema + dictionary + DOI) and reduces/denies credits when FAIR elements are missing. Issuances record the FAIR rubric version and score in the ledger, ensuring traceable rationale for awards.

We operationalize fairness & bias through required documentation and lightweight diagnostics for datasets and models submitted in Epic 2. Contributors attach a Datasheet/Model Card that states population coverage, collection context, known limitations, sensitive attributes handling, and intended use. Where applicable, the pipeline computes basic imbalance summaries (e.g., class distributions) and flags obvious risks for curator review. Submissions that include bias-aware diagnostics, mitigation notes, and reproducible evaluation notebooks earn bonus credits; omissions trigger returns or reduced awards. For synthetic datasets, contributors provide a

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Synthetic Data Card noting privacy leakage tests and utility metrics; provenance tags and risk profiles are mandatory.

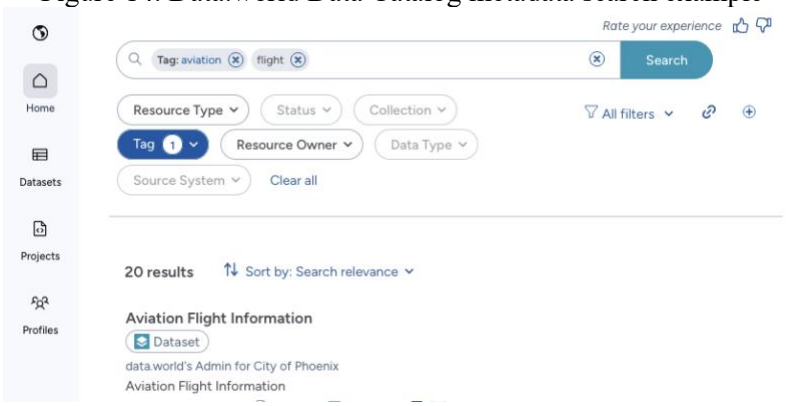
Spending in Epic 2 focuses on premium datasets/models and rate-limited APIs. We surface responsible-use context at the point of spend: the portal displays the artifact's license, known limitations, and any flagged fairness notes before confirmation. The spend record in the ledger captures the artifact ID, policy version, and any responsible-use attestations the user acknowledges, maintaining transparency without blocking legitimate access.

Finally, we make FAIR and fairness visible and improvable. The portal shows a FAIR & Responsible Use badge on approved contributions with a drill-down of passed checks (metadata completeness, license, provenance, bias docs). Organization dashboards roll up FAIR scores and responsible-AI KPIs (e.g., % with Datasheets/Model Cards; % with imbalance summaries; DLP issues resolved). All rubric versions, reviewer decisions, contributor attestations, and issuance rationales are immutably referenced in the credits ledger, closing the loop between governance and incentives in Epic 2.

Dataset Discovery and Search – Requirement 8.3

Name	#	Epic 1 (Data Catalog)	Epic 2 (Credits & Upload)	Epic 3 (Synthetic Data)
Dataset Discovery and Search	8.3	100%		

Figure 14. Data.world Data Catalog metadata search example



As described in *Dataset Directory – Requirements 5.1, 5.2*, users can search over multiple customizable facets. Full-text search is not available and not recommended due to the size of the data we anticipate surfacing. However, we can create a full-text search solution if this is an essential component of what the DCC is looking for. Users can natively search across metadata (including column metadata and descriptions, see figure 14). All search functionality is also natively available with the REST API via the "search" endpoint: https://developer.data.world/reference/searchresources1.

Access Control and Governance – Requirement 8.4

Name	#	Epic 1 (Data Catalog)	Epic 2 (Credits & Upload)	Epic 3 (Synthetic Data)
Access Control and Governance	8.4	100%		

Use or disclosure of data contained on this page is subject to the restriction on the title page.

All users of the platform and their actions are governed by role-based access controls, including the ability to define who can view edit, publish, or delete data both platform-wide and on individual or groups of datasets. More details on specific roles can be found in the *Access Controls – Requirement 13.1* section below.

Public access will be limited to the landing page and public content linked on the landing page. Once a user is registered, they will have access to the data catalog and companion sites (such as the credits system and synthetic data platform) and are limited to approved content according to their roles. Administrators of the platform will have elevated privileges according to their roles (more specific can be found in the *Access Controls – Requirement 13.1* section below).

Data.world offers a comprehensive and curated collection of logging data referred to as "base platform data" with deep, actionable insights into platform usage, user activity, and resource management. The following data are included in this collection:

- **Audit events** – Audit events allow administrators to monitor the actions performed by all the users in the data.world application through the UI or while using API. The audit log reporting functionality enhances the accountability of actions in the application. Administrators can track the actions taken by users in the application and find the root cause of issues by identifying the resources on which the action was performed and who performed the action.
- **Platform analytics** – A collection of resources that track and analyze various aspects of catalog usage and user behavior. It includes daily summaries of dataset interactions, user activity, API usage, and metadata management. Provides comprehensive insights into platform engagement, resource utilization, and customer trends. These tables are essential for monitoring the overall health and performance of the platform.
- **Platform events** – The platform events capture how users interact and engage with the platform. The events provide critical insights into user behaviors, such as identity management, resource engagement, content downloads, and navigational patterns, enabling a comprehensive understanding of how users interact with the platform.
- **Platform records** – Platform records contain records of user activities and interactions across various platforms, including transaction history, preferences, and engagement metrics. It also stores information on user profiles, such as account details and demographic information. This provides a comprehensive view of user behavior and preferences.

All the above data is readily queryable within the back-end Snowflake platform. All custom development work will post events into AWS CloudTrail logs. These logs will also be synchronized into Snowflake tables for centralized consumption of log information.

Dataset Publishing & Versioning – Requirement 8.5

Already discussed in *Dataset Selection & Upload Process – Requirements 5.4, 5.6, 8.5, 9.1, 9.2* section above. This section is here to aid in navigation of this proposal.

Interoperability & Integration – Requirement 8.6

Name	#	Epic 1 (Data Catalog)	Epic 2 (Credits & Upload)	Epic 3 (Synthetic Data)
Interoperability and Integration	8.6	100%		

Use or disclosure of data contained on this page is subject to the restriction on the title page.

All datasets sourced into the data.world catalog are downloadable into machine-readable formats or used directly via the REST API or one of the following pre-built SDKs:

- Python SDK (including seamless integration with Jupyter notebooks)
- R SDK (including seamless integration with R Studio)
- Go SDK
- JDBC Driver

The platform enhances datasets with metadata using various RDF ontologies, enabling structured data querying in accordance with Linked Data. The ontologies used include RDF/RDFS for class and property definitions, OWL for property characteristics, and more. Data.world automatically generates metadata triples during data ingestion, representing table and column definitions, dataset descriptions, and statistical summaries.
As described in the *Metadata Capabilities and Management – Requirements 8.1, 8.2* section above, we adopt a "FAIR-by-default" metadata profile that blends descriptive, technical and provenance fields aligned to common schemas (Dublin Core, Schema.org, DCAT) so records are discoverable, interoperable, and consistently documented across hosted and linked assets. In practice, the submission form and curator tools enforce these fields (title, abstract, keywords, license; file formats/schemas/dates; source and processing lineage), and the catalog exposes them via APIs for programmatic reuse. This directly implements Descriptive metadata, technical metadata, provenance, and explicit support for metadata standards.

As described in the *Metadata Capabilities and Management – Requirements 8.1, 8.2* section above, we include a standard citation block on each dataset page (creator, title, year, version, publisher/host, URL) and support persistent identifiers (e.g. DOIs via DataCite) for stable, hosted releases. The catalog tracks usage and citation linkages over time, satisfying the RFP's emphasis on citation tracking and DOI readiness. Practically upon approval the workflow can mint or record a DOI and write it into the data sets metadata, while the catalog maintains download/access metrics and citations.

Data Usage and Metrics – Requirement 8.7
Already discussed in the *Admin Dashboard – Requirements 5.8, 8.7* section above. This section is here to aid in navigation of this proposal.

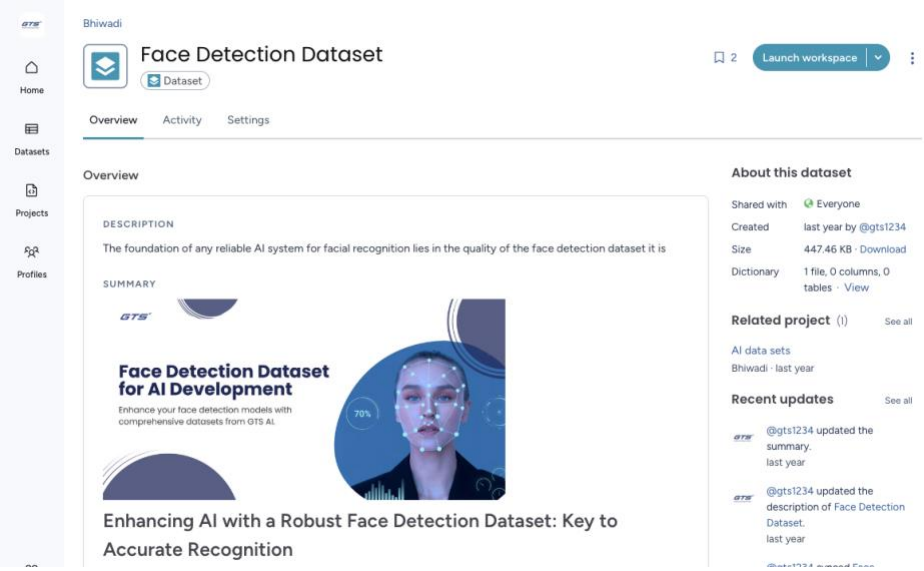
Collaboration and Community Features – Requirement 8.8
Already discussed in the *Community Dataset Submissions – Requirements 5.7, 8.8* section above. This section is here to aid in navigation of this proposal.

Data Previews and Exploration – Requirement 8.9

Name	#	Epic 1 (Data Catalog)	Epic 2 (Credits & Upload)	Epic 3 (Synthetic Data)
Data Previews and Exploration	8.9	100%		

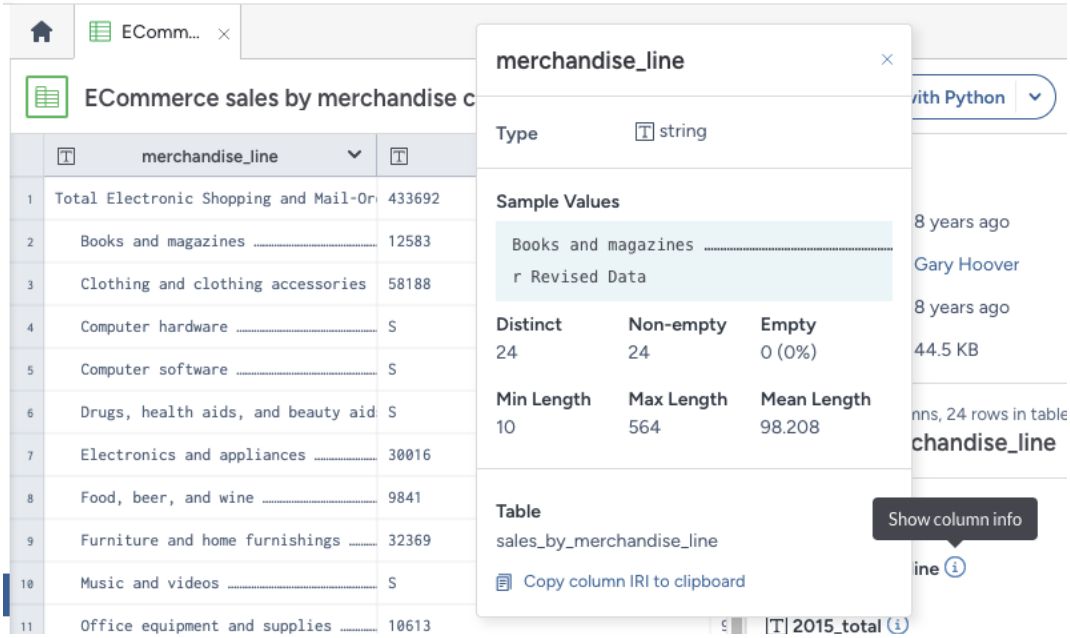
Use or disclosure of data contained on this page is subject to the restriction on the title page.

Figure 15. Data.world Data Catalog Data Preview example



As demonstrated in the previous sections, in-browser previews are already available for CSV or JSON files without downloading, see figure 15. Sample images can also be included as part of the description or as standalone files that can be previewed in the data catalog.

Figure 16. Data.world Data Catalog summary statistic example



Each data field has a "Column info" view that provides summary statistics including a count of distinct values, percentage empty, and min/max/mean length, see Figure 16. The data.world catalog offers the ability to integrate with visualization tools including Tableau, Looker, R Studio, Jupyter notebooks, or the native Chart Builder tool.

Example Data Management and Catalog Capabilities for Consideration – Requirement 8.10

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Name	#	Epic 1 (Data Catalog)	Epic 2 (Credits & Upload)	Epic 3 (Synthetic Data)
Example Data Management and Catalog Capabilities for Consideration	8.10	100%		

While our team evaluated both open source and commercial options for the data management and catalog portion of this RFP, we were convinced data.world provided the most value compared to other options for the following reasons:

1. **Automation & Long-Term Maintenance** – With data.world, automation is a first-class citizen, particularly with its recent integration with ServiceNow. As frequent consumers of open data platforms, we often find that when they first launch, they are easy-to-use, and the datasets are accessible. However, over time these tools degrade in utility and links are no longer accessible, leading to a frustrating user experience. As a part of our proposal, we will fully leverage data.world's automation capabilities to ensure the DCC remains high quality post-launch with minimal resources required for ongoing maintenance.

2. **Open Standards, Commercial Support** – data.world was legally established as a public benefit corporation / certified B Corporation, which means its corporate mission was explicitly oriented toward serving the public good. Because of this charter, data.world prioritized long-term community value and openness alongside commercial viability. Over the past decade, it has grown into the world's largest open data community, with millions of contributors, dozens of institutional partnerships, and thousands of active datasets shared publicly. Because the open community is core to the platform's identity, data.world is architected with open standards and extensibility in mind. This provides all of the benefits of the open source tools, while the commercial aspect ensures that the DCC benefits from active development, frequent feature releases, and dedicated support.

3. **Feature-Rich, User Friendly** – The data.world catalog is widely praised for its user-friendly interface and workflow design, which significantly reduces friction and accelerates onboarding of new users. Its collaborative features (such as easy dataset annotation, social functionality, and business glossary support) make it a preferred choice for teams needing seamless data discovery and shared context.

Criteria for Inclusion & Suggested Initial Datasets – Requirements 9.1, 9.2

Already discussed in *Dataset Selection & Upload Process – Requirements 5.4, 5.6, 8.5, 9.1, 9.2* section above. This section is here to aid in navigation of this proposal.

Synthetic Data Generation Capability – Requirements 10.1 – 10.3

Already discussed in *Synthetic Data Generation Capability – Requirements 5.5, 10.1-10.3* section above. This section is here to aid in navigation of this proposal.

AI Fairness and Bias Capability – Requirements 11.1

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Name	#	Epic 1 (Data Catalog)	Epic 2 (Credits & Upload)	Epic 3 (Synthetic Data)
Purpose	11.1	100%		

FAIR by design, the hub implements the FAIR principles –Findable, Accessible, Interoperable and Reusable through a standards-based catalog and a policy-backed publishing workflow. Every dataset is described with rich metadata (DCAT/ Schema.org /Dublin Core), controlled vocabularies, and keywords to support powerful faceted and full-text search. On approval, catalog entries and previews/APIs go live and the portal re-indexes automatically so new records are immediately discoverable. This ensures findability and reuse without release steps.

Contributors enter through a guided submission that captures descriptive, technical, and provenance metadata along with license and access tier. Curators standardize tags, validate formats, and enable previews; approvers verify FAIR completeness and policy alignment before publication. The admin console provides moderation queues, role management, link-health monitoring, and a full audit trail of decisions.

To align with the RFP's oversight objectives, the pipeline incorporates a targeted fairness screen via AIF360 for submissions likely to inform modeling or decision-making, especially datasets containing protected attributes or labeled outcomes. Curators can trigger a containerized check that computes core metric (eg. Disparate impact, statistical parity difference, equal opportunity) and attaches a human-readable "Fairness Card", JSON metrics and an HTML report to the catalog record. Approvals are gated by configurable thresholds, and approvers may require mitigation or add justification for borderline cases. The dataset page surfaces a simple badge and headline metrics for public transparency, with accessible plain-language summaries.

Governance Requirements – Requirements 12.0-12.5

Name	#	Epic 1 (Data Catalog)	Epic 2 (Credits & Upload)	Epic 3 (Synthetic Data)
Governance Requirements	12.0	50%	25%	25%
Core Capabilities	12.1	50%	25%	25%
Operational Requirements	12.2	50%	25%	25%
Governance and Legal Compliance	12.3	50%	25%	25%
Technical Integration	12.4	50%	25%	25%
Monitoring and Reporting	12.5	50%	25%	25%

Our governance model operationalizes transparency, accountability, and oversight as a workflow, not a one-time review. It is centered on community submissions, moderation approvals, and auditability with clear decision

Use or disclosure of data contained on this page is subject to the restriction on the title page.

rights and role separation in an admin console. Submissions land in a curated queue; reviewers standardize metadata and enforce policy; approve finalize outcomes. All steps are logged with rational and are visible for program oversight.

Role and decision rights are handled as follows: We separate duties across Submitter -> Curator -> Approver - > Publisher, with delegated curator roles to support multi-tenant participation. Approvals and rejections are captured with justification; escalations and waivers follow defined pathways. This aligns with the RFPs expectations for moderation, role management, and administrative control.

Policy gates and exceptions are handled via governance checks. Governance checks which encompasses data classification/tiering, license compliance, privacy/PII flags, prerequisite metadata completeness are encoded as gate conditions in the approval flow. Gate outcomes are Pass /Borderline (mitigation note requires) / Fail (changes requested). Exception approvals require explicit justification and become part of the audit trail. (This complements but does not duplicate the fairness screen, which is treated as one gate signal among several) Operational controls and transparency are fulfilled through the admin dashboard. The admin dashboard provides: submission aging/SLA views, link-health monitoring for referenced datasets, role administration, and a searchable log of decisions. This meets the RFP's requirements for operational administration and ongoing moderation, ensuring community visibility and accountability.

Governance operates under program security controls such as SSO/MFA, RBAC, dataset tiering, audit logging and applies them consistently to submission, review, and publication processes. This ensures oversight decisions are enforceable at access time and traceable after the fact.

For collaboration and feedback loops, the catalog entries support discussion and annotations and usage metrics to inform curation priorities and measure impact over time providing a transparent feedback loop between contributors, curators, and consumers.

In Epic 2, we extend the governance model by operating the Credits & Contributions program under a documented governance charter owned by the Program Owner (MassTech) and executed by a Governance Team. The charter defines scope (what can earn/spend credits), decision rights, policy lifecycles, and evidence requirements. Policies (issuance rules, price tables, caps, FAIR/fairness rubrics) are versioned, change-controlled, and time-boxed; each issuance or debit records the policy version in the immutable ledger so any outcome is reproducible. A monthly operating committee reviews posture (balances, burn, exceptions) and a quarterly steering committee ratifies policy changes.

We make the program explainable to users and auditable to oversight. The portal publishes plain-language policy pages, current price tables, FAIR/fairness rubrics, and "how credits work" guides. Every transaction (earn/spend/adjust/reverse) is logged with actor, organization, artifact IDs, timestamps, and policy version; users see their full history and organizations see roll-ups. Administrators export signed reports from the ledger and workflow system for internal/external review. Public artifacts (e.g., rubrics) are change-logged with effective dates so stakeholders can see what changed and when.
Epic 2 introduces credits-aware roles and strict segregation of duties. Submitters contribute; Curators/Reviewers approve or return; Credits Admins post compensating transactions and set caps; Policy Admins edit price tables and multipliers; Auditors have read-only access; Org Admins manage team visibility; Break-glass Admins are time-boxed for incidents only. No single role can approve a submission **and** issue credits for it, or set a price **and** spend under that price. Elevated actions require step-up MFA and a maker–checker workflow; quarterly access recertifications remove stale privileges.

Submissions must include machine-readable licenses (SPDX), provenance, usage notes, and where applicable consent or data-sharing attestations. The workflow enforces license compatibility and provenance checks, DLP/PII screening, and metadata completeness before approval. Sensitive uploads are quarantined until a curator

Use or disclosure of data contained on this page is subject to the restriction on the title page.

disposition is recorded. We retain all approvals, attestations, and exceptions with links to the source artifacts and policy versions. Right-to-erasure requests trigger pseudonymization in workflow records while preserving the ledger's integrity via compensating entries (no destructive edits to audit logs).

We continuously monitor for governance risks: burst adjustments, abnormal spend, repeated submission failures, policy drift, and DLP hits. Anomaly detections open tracked cases; sanctions (temporary caps, account holds) are applied per policy with notice to the organization. A runbook defines severity classes, containment steps, communications, and recovery.

Contributors can appeal denials or debit disputes through a transparent, tracked process. Investigators review evidence (ledger entries, workflow decisions, rubrics) and either uphold the outcome or post a compensating transaction with rationale. We publish program KPIs e.g., approval rates, time-to-decision, FAIR score distribution, reversal rate and solicit feedback through structured surveys and office hours. On a quarterly cadence, the Governance Team updates rubrics and policies (e.g., FAIR criteria, incentive weights), announces changes ahead of enforcement, and advances the policy version so historical transactions remain comparable and auditable.

Education & Capacity Building – Requirement 12.6

Name	#	Epic 1 (Data Catalog)	Epic 2 (Credits & Upload)	Epic 3 (Synthetic Data)
Education and Capacity Building	12.6	40%	30%	30%

1. Overview

The success of the Commonwealth's Data.world catalog implementation depends on more than just technical deployment; it requires a structured training and adoption strategy. Our approach ensures that employees, administrators, and stakeholders are fully enabled to use, manage, and promote the catalog. Training will emphasize both **practical navigation skills** and **strategic benefits**, fostering widespread adoption and sustainable value.

2. Training Strategy

2.1 User Training – General Workforce

Archetype's training program is designed to ensure all Commonwealth employees can confidently navigate and use the Data Commons Collaborative (DCC) platform to its full potential. Training aligns directly with RFP functional requirements, including **dataset search and discovery (Section 5.2–5.3)**, **metadata exploration and FAIR compliance (Section 8.2)**, and **synthetic data awareness (Section 10.1–10.3)**.

- **Navigation & Access:** Conduct role-based sessions that train users to log in via Entra ID SSO, navigate the catalog interface, and efficiently search and filter datasets by domain, format, or metadata attributes.
- **Use and Benefits:** Deliver scenario-based instruction demonstrating how catalog features—such as dataset discovery, synthetic data generation, and FAIR metadata—enhance collaboration and decision-making.
- **Explorer Features:** Provide interactive walkthroughs of advanced capabilities, including lineage visualization, metadata search, dataset previews, and filtering by provenance or version.
- **Access Tools:** Educate users on data access mechanisms (APIs, BI connectors, or Snowflake secure shares), showing how to responsibly use datasets within approved analytics tools.

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Training materials will leverage **data.world's existing user guides, tutorials, and documentation** as a foundation—adapted to Commonwealth-specific use cases, governance standards, and data types. This approach accelerates adoption and ensures users benefit from proven instructional content while maintaining compliance with DCC policies.

2.2 Administrator Training – Stewards & Technical Leads

Administrator and steward training focuses on managing catalog integrity, governance workflows, and data quality assurance as outlined in **Sections 8.4 (Access Control), 8.5 (Publishing and Versioning), and 12.0 (Governance Requirements)**.

- **System Administration:** Conduct deep-dive sessions covering catalog permissions, metadata curation, access control tiers, and governance enforcement through ServiceNow integration.
- **Community Engagement:** Equip administrators to manage community submissions, moderate datasets, and apply metadata validation workflows to maintain catalog trust and consistency.
- **Sustainability:** Train stewards on monitoring metadata completeness, dataset versioning, and usage analytics to ensure ongoing catalog health.

All administrator sessions will include references to **data.world's open training resources and admin documentation**, customized to reflect DCC metadata models, lineage policies, and governance framework alignment.

3. Training Delivery & Documentation

- **Launch Communications:** Announce the DCC launch with communications that highlight catalog benefits, promote Massachusetts datasets, and direct users to training resources.
- **Documentation:** Provide interactive user guides, quick reference cards, and short video tutorials built from **data.world's documentation framework**, ensuring consistent messaging and usability.
- **Accessibility & Responsiveness:** Ensure all materials meet **WCAG and Section 508 standards**, with responsive design for use across desktop, tablet, and mobile environments.

4. Adoption and Engagement

- **Showcase Massachusetts Datasets:** Use state-owned datasets (e.g., public health, housing, transportation) as live examples in demos and exercises to make training immediately relevant.
- **Community Building:** Promote engagement through forums, "power user" spotlights, and cross-agency collaboration sessions.
- **Feedback Loops:** Implement continuous improvement through feedback surveys and analytics on training effectiveness, catalog usage, and feature adoption.

5. Implementation Phases

Phase	Focus	Key Activities
Pre-Launch (Awareness & Preparation)	Build readiness	Develop training materials, adapt data.world guides, and prepare launch communications. Identify and train administrators and stewards.
Launch (Enablement & Adoption)	Enable early users	Deliver user and admin training sessions focused on navigation, FAIR metadata, and synthetic data examples. Showcase Massachusetts datasets and launch community engagement forums.
Post-Launch (Sustainment & Growth)	Drive adoption	Offer refresher and onboarding sessions, advance admin governance skills, and track adoption metrics to refine content and delivery.

Use or disclosure of data contained on this page is subject to the restriction on the title page.

6. Outcome

This structured approach ensures that the data.world catalog is not only deployed but **actively adopted and sustained** across the Commonwealth. Users will gain the confidence to navigate and leverage the catalog effectively, administrators will be equipped to manage and grow the community, and the Commonwealth will achieve its vision of improved data discovery, governance, and value realization at scale.

Access Controls – Requirement 13.1

Name	#	Epic 1 (Data Catalog)	Epic 2 (Credits & Upload)	Epic 3 (Synthetic Data)
Access Controls	13.1	60%	20%	20%

Our role-based access control (RBAC) model applies consistently across the portal, APIs, catalog, hosted storage/compute, and workflow tools. We enforce least privilege, separation of duties, and defense in depth. A users identity comes from SSO/MFA and their roles come from groups/claims while enforcement happens at every layer of the architecture including UI, APIs, catalog objects, dataset/table.

Every dataset carries an access tier in its metadata Public, Restricted, Internal plus license/usage constraints. That tier is evaluated at request time so the same rules apply whether access happens via website, the catalog API, or direct data connections. Public resources are open, restricted resources require an approved entitlement while Internal maps to designated program roles. This also lets us surface clear UX cues and keep audit trails tied to the dataset record.

Users authenticate through our IDP with MFA. Group membership maps to program roles such as Viewer, Contributor, Curator, Approver, Admin. Those role claims are consumed by the portal, ServiceNow, data.world, and Snowflake so the user's effective privileges are consistent everywhere. Approvals and role changes are logged for audit in the admin console.

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Enforcement by Layer
Portal and APIs
The web tier checks JWT role claims before rendering content or issuing downloads. The API Gateway and Lambda will be leveraged as part of the AWS Amplify framework for enforcing scopes and rate limits; all access decisions and denials are logged.
Data Catalog (data.world)
Dataset/project permissions mirror program roles (read, write, admin). On publish, the workflow sets visibility according to tier and writes role binding; collaboration features (comments, metrics) respect the same permissions.
Hosted data platform (Snowflake)
Snowflake roles align to program roles; data is exposed through schemas, secure views, and (where needed) row/column policies that match the dataset tier. Token-based connections from the portal catalog inherit the requester's effective role.
Workflow & Admin (ServiceNow)
Submission, curation, and approval actions are restricted by role; the Admin dashboard provides moderation queues, role management, and a full audit trail of decisions and access changes.

We use time-boxed entitlements for elevated access (auto-expire unless renewed), service accounts with minimal scopes for automation, and environment segregation with distinct role sets. All sensitive operations emit auditable events retained per policy and metrics and logs support security reviews and incident response.

In Epic 2, we extend the platform's RBAC from "catalog-first" to "credits-and-contributions–aware" controls. Identities continue to federate from the Commonwealth/MTC IdP, but tokens now carry scoped claims for organization, role, and capabilities specific to Credits and Submissions for example: credits:issue, credits:spend, submission:approve. Access decisions are enforced at the API Gateway (JWT authorizer + usage plans) and again inside services, ensuring least privilege at the edge and in the business layer. All ledger writes (earn/spend/adjust/reverse) are executed exclusively by service principals; humans never write directly to the ledger, eliminating bypass risk.

We introduce segregation of duties so no single persona can both approve a contribution and issue credits for it, or change pricing rules and spend credits under those rules. Elevated actions require step-up MFA and a maker–checker workflow. RBAC scopes are tenant-aware: users operate within their organization by default; cross-org actions for example program audits are limited to designated roles. Administrative scopes are time-boxed and must be re-granted for each change window; all scope changes are logged and included in quarterly access recertification.

- Submitter (Org-scoped): create/edit submissions; view own credit balance & history; no authority to approve, issue, or reverse credits.
- Curator/Reviewer (Delegated): triage and approve/return submissions; cannot issue credits; cannot change credit policies or prices.

Use or disclosure of data contained on this page is subject to the restriction on the title page.

- Credits Admin (Program-scoped): manage issuance adjustments and refunds via compensating transactions; set org/user credit caps; cannot approve submissions; cannot modify price tables.
- Policy Admin (Program-scoped): edit price tables, multipliers, and incentive windows; cannot issue or spend credits; cannot approve submissions.
- Org Admin (Org-scoped): view org-level balances and transactions; request cap changes; cannot set global policies; cannot issue credits.
- Researcher/User (Org-scoped): spend credits on permitted resources (datasets, APIs, MGHPCC/AICR jobs) subject to caps and throttles; read-only transaction history.
- Auditor (Read-only, Program-scoped): access full ledger and policy history; no write permissions anywhere.
- Break-glass Admin (Program-scoped, time-boxed): incident-only powers to unblock stuck transactions; every action requires ticket linkage and generates auto-notifications.

We enforce resource-level permissions for the new objects introduced in Epic 2: Submission, ReviewTask, CreditPolicy, PriceTable, CreditCap, and LedgerTransaction. Each API route declares the exact scopes required. Rate limits and per-scope quotas prevent abuse . Finally, we codify access lifecycle: joiner–mover–leaver flows automatically grant, update, or revoke Epic 2 scopes; quarterly role recertifications and access anomaly reports are delivered to Program Owners for sign-off.

In Epic 3, we further extend the RBAC model to introduce the following roles which govern usage and administration of the synthetic data generation (SDG) capability.

- No SDG Access – can request access access through web portal
- SDG Contributor – can initiate synthetic data generation requests through portal, can submit synthetic data outputs for usage in data catalog, has access to synthetic data generation jobs they initiate
- SDG Access Administrator – can approve access requests
- SDG Auditor – has access to all synthetic data generation jobs metadata
- SDG Contribution Approver – has access to synthetic data jobs submitted for approval
- SDG Break-Glass Admin – has access to all synthetic data generation jobs and the underlying data. Only to be used in cases where terms are violated.

Data Protection – Requirement 13.2

Name	#	Epic 1 (Data Catalog)	Epic 2 (Credits & Upload)	Epic 3 (Synthetic Data)
Data Protection	13.2	60%	20%	20%

All interactions between all systems in this proposal are encrypted with TLS 1.2+ encryption while in transit. All information stored in data.world, AWS S3, and Snowflake are encrypted at rest, leveraging the keys provided these platform providers. If needed, customer-managed keys can be leveraged for further control of encryption standards.

Our deployment leverages AWS CloudWatch to enable intrusion detection and monitoring by collecting and aggregating security logs from hosts and applications, applying metric filters to spot suspicious patterns like failed logins, and generating automated alerts to notify or trigger remediation actions. CloudWatch's built-in anomaly detection uses machine learning algorithms to baseline metrics and flag unusual activity that may indicate security

Use or disclosure of data contained on this page is subject to the restriction on the title page.

incidents. For a robust defense, it integrates with services like GuardDuty and Security Hub and can be combined with network and service-based protections to support a layered security strategy.

Our team leverages a zero-trust compliance framework and analysis accelerator that we use with all of our customers to conduct regular security reviews. For penetration testing, we use pentest-tools' Website Scanner and API Vulnerability scanner to ensure that the publicly accessible sites are not vulnerable to any known exploits. Additionally, we leverage AWS Inspector to scan for vulnerabilities within our cloud resources.

In Epic 2, we extend the platform's data protection posture to cover new data classes introduced in Epic 2: Submission records and attachments, ReviewTask artifacts, CreditPolicy/PriceTable configurations, organization/user credit caps, and ledger transactions. Each class is cataloged with classification (Public, Internal, Restricted) and purpose binding. By default, submissions and ledger data are Internal; attachments may be auto-escalated to Restricted if content scanners detect sensitive elements. We enforce data minimization: the ledger stores pseudonymous references (user/org IDs) and purpose codes rather than free-text PII; descriptive context lives in workflow records with RBAC.

All Epic-2 data is encrypted in transit (TLS 1.2+) and at rest: QLDB (credits ledger), DynamoDB store (workflow state), S3 (attachments/exports) with customer-managed KMS keys, per-environment key separation, rotation, and automatic re-encryption on policy change. Secrets (API keys, webhooks, broker creds) are kept in AWS Secrets Manager, rotated on a schedule and on role changes; step-up MFA is required for any secret retrieval via console.

Credits and Submission services run in segmented VPCs, with private endpoints, security groups, and per-service IAM roles scoped to the minimum read/write set (e.g., only the Credits Service role can write to QLDB journal tables). Administrative actions (grants, reversals, policy edits) require maker–checker approval and step-up MFA, and produce tamper-evident audit entries that link back to the initiating ticket. Humans never write directly to the ledger; only the service performs ACID mutations with idempotency keys. Cross-org data access is disabled by default; org-scoped reads are enforced in the data layer (row-level filters) and at the API.

We deploy two complementary controls: (1) an IDPS layer (WAF, Guard Duty, Network Firewall, API Gateway throttling) that protects the platform from hostile traffic and anomalous behavior; and (2) an inbound content–scanning pipeline that validates every submission before it can be curated or earn credits. The content pipeline is dataset-oriented: it enforces a strict MIME/format allowlist (CSV/JSON/Parquet/Avro, plus approved docs), verifies magic bytes to prevent spoofed extensions, and performs schema sanity checks . It runs malware/active-content detection, applies license/provenance checks, and executes DLP/PII inspection to quarantine risky uploads. Failures route to a curator for disposition; approved objects are promoted with scan attestations and immutable provenance. Together, the IDPS and content scanning ensure we control both how data arrives and what can enter the catalog.

We centralize logs (API Gateway, Lambda, QLDB, workflow engine, scanners) with immutable storage and **≥3-year** retention for governance. Anomaly rules detect high-risk patterns (burst adjustments, cap changes, failed MFA, repeated scanner hits). Daily **r**econciliations validate that ledger balances match issuance–debit deltas; variances open tickets automatically. Backups and journal exports are written to versioned, WORM-locked S3 buckets with lifecycle policies; restores and disaster-recovery procedures are rehearsed at least annually.

We apply retention schedules by class: submissions and decisions per records policy; scanner artifacts 90–180 days; operational telemetry per security policy; ledger indefinite or per statutory requirement. Right-to-erasure requests never delete ledger entries; we post compensating entries and pseudonymize subject identifiers, preserving audit integrity. Public dashboards use aggregated, k-anonymous metrics; no raw PII is exposed.

We run pre-GA penetration tests on Credits, Submissions, and admin surfaces; remediate before release, most likely leveraging a third-party firm. Containers/functions receive SCA/SAST in CI; dependencies are pinned and auto patched. We maintain an incident response runbook (classification, containment, forensics, notification).

Use or disclosure of data contained on this page is subject to the restriction on the title page.

In Epic 3, we ensure that all previously established access control guidelines are maintained as we build out additional infrastructure and web features.

Governance Policies and Ethical Guidelines – Requirement 14.0

Name	#	Epic 1 (Data Catalog)	Epic 2 (Credits & Upload)	Epic 3 (Synthetic Data)
Guiding Principles	14.0	100%		

Throughout this response, we have provided detailed examples of how our proposed solution embodies the guiding principles laid out in this requirement. Below is a summarization of how these principles are addressed throughout the proposal.

Transparency

- The proposal commits to publishing comprehensive documentation on all data sources utilized, detailing both the origins and attributes of datasets ingested.
- Transformation processes (including data cleaning, standardization, and any synthetic data generation) are described in accessible platform documentation, available to stakeholders and the public where relevant.
- Clear audit trails for data lineage and transformation are maintained, enabling users to trace changes from original raw sources through intermediate steps to final data products.

Accountability

- Our solution includes robust logging of data access activities, specifying who accessed what data, when, and for what purpose.
- Role-based access controls are deployed, ensuring that only authorized personnel can interact with sensitive datasets.
- Usage monitoring policies are defined, with regular reviews and alerts for any anomalous or potentially irresponsible activity.
- A data governance framework assigns responsibilities for oversight and escalation, enforcing adherence to responsible data use policies.

Equity

- The proposal outlines protocols for evaluating datasets before inclusion, with statistical and qualitative reviews specifically designed to identify and mitigate bias.
- There are minimum standards for representativeness, requiring the documentation and review of demographic parameters and sampling methods.
- Automated and manual mechanisms flag datasets or transformations that risk exacerbating bias or marginalizing underrepresented groups.
- Stakeholder engagement processes (such as reviews and feedback loops with domain experts and community advocates) are implemented to continually assess equity risks.

Consent and Compliance

- All datasets are sourced and processed in accordance with applicable privacy and data protection laws (such as HIPAA, GDPR where relevant), with explicit verification steps for consent and licensing documentation.
- An ongoing compliance audit system forms part of the data onboarding process, ensuring any new or updated dataset meets legal requirements before integration.
- The proposal stresses the importance of data minimization and de-identification, especially for sensitive or personal data, to minimize regulatory risk.
- Regular training for staff and contributors reinforces best practices on consent, lawful data use, and regulatory obligations.

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Future Considerations

With a Massachusetts-based company as the lead for this proposal and a combined team of data, AI, technology, and governance enthusiasts, we are very excited for the prospects of this project. We are AI optimists that recognize the importance of a public-private-academic partnership in ensuring that our society can benefit from the benefits of leveraging AI, while simultaneously putting in common sense guardrails to ensure that these tools are used safely and in equitable manner. In our proposal, we have carefully thought through ways to meet the requirements outlined in the RFP, while ensuring the platform is easy to both maintain and extend.

In the spirit of extensibility, we wanted to outline a few ways this project can be potentially expanded in the future and how the tools and processes we put in place for this MVP aid in enabling these potential use cases.

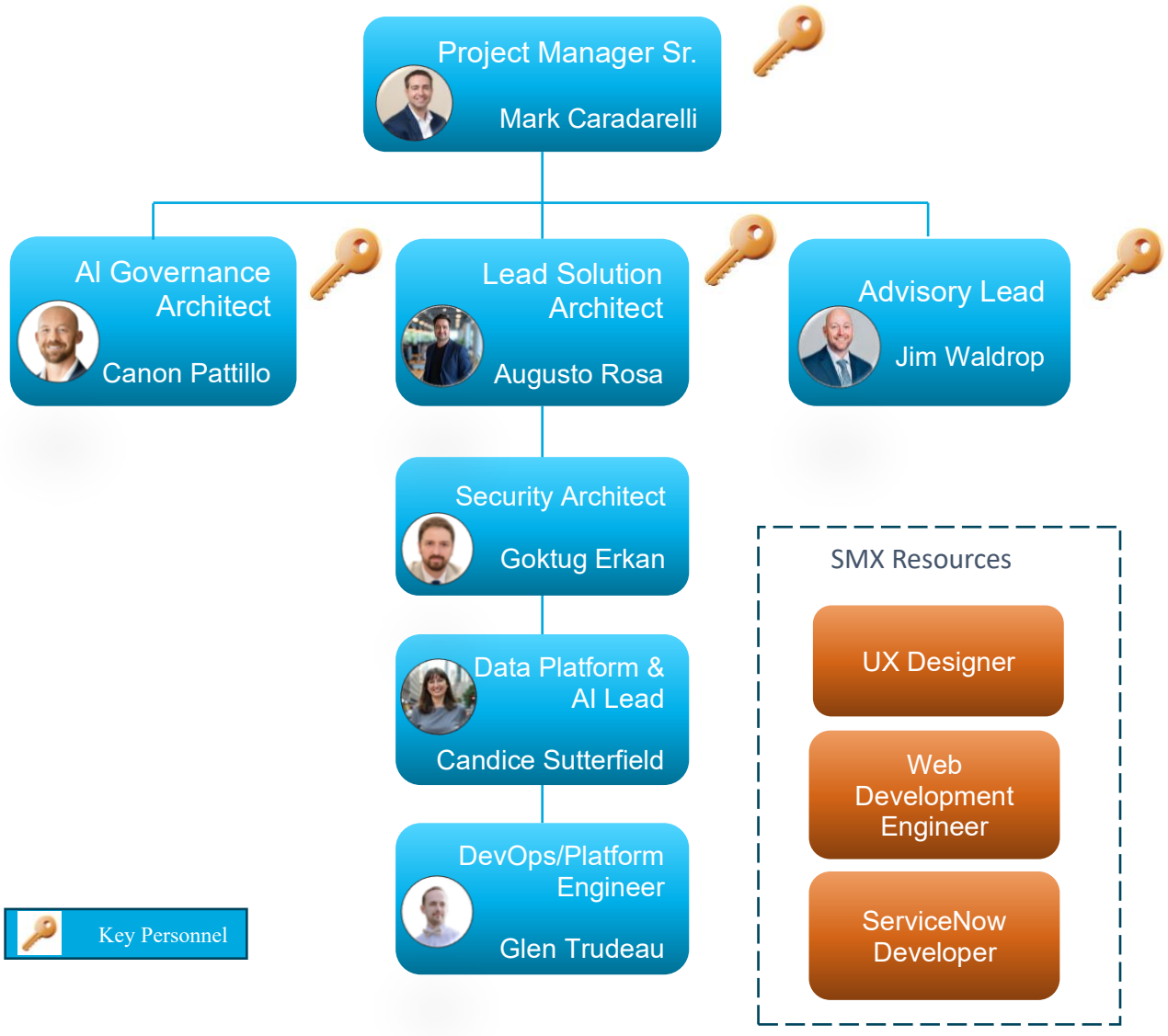
- **Semantic Layer for Natural Language Querying** – Implement an enterprise-grade semantic layer within the Data Commons Collaborative, connecting to both structured and unstructured data sources using the existing ontologies and knowledge graphs within data.world to support a broader tier of users that can benefit from the datasets managed by the DCC.
- **Extending Synthcity with Additional Model Modules** – Collaborate with Synthcity open-source community and the DCC community to design and integrate new modules; specifically supporting state of the art image data generation or video data generation.
- **Model Training Capabilities** – Integrate dedicated pipelines for machine learning and AI model development directly into the AICR's HPC infrastructure in a similar manner to the synthetic data generation mechanisms.
- **Integrated Data Visualization and Storytelling Tools** – Leverage data.world integrations to allow for the DCC community to incorporate data visualizations as catalog entities with lineage to DCC data sources maintained. This fosters broader community engagement and helps illustrate the impact of discoveries to funders and policymakers.
- **MCP Repository for Agentic Workflows** – Create an MCP server repository to collect tools and context for help in developing agentic workflows. Develop MCP servers that expose tools and datasets within the DCC catalog for use in LLM-based use cases.

4.0. Key Personnel

Archetype's management structure is designed to ensure clear accountability, cross-functional coordination, and rapid decision-making throughout the AI Hub Data Commons program. Led by **Project Manager Mark Cardarelli**, the team integrates governance, advisory, and technical leadership into a unified delivery framework. **Advisory Lead Jim Waldrop** drives statewide training, change management, client value realization, and governance framework development, ensuring alignment with EOTSS and ODIA standards. **Lead Solution Architect Augusto Rosa** provides technical oversight across all Epics, supported by **AI Governance Architect Canon Pattillo**, who ensures compliance with metadata, fairness, and policy requirements. The team is further strengthened by **SMX specialists** in UX, web development, and ServiceNow automation, along with **security, data platform, and DevOps engineers** who maintain operational resilience and compliance. Together, this structure provides the Commonwealth with both the strategic vision and technical discipline required to deliver a secure, FAIR-compliant, and scalable data commons.

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Figure 17: Organizational Chart



Tier	Role	Primary Focus
Program Leadership	Project Manager	Day-to-day management, schedule, coordination
Strategic & Transformational Leadership	Advisory Lead	Drive client value & outcome alignment, lead use case development & prioritization, oversee governance framework & policy design, executive stakeholder liaison & communications
Governance & Compliance	AI Governance Architect	Develop governance framework, AI ethics, policies
	Security Architect	Security design, compliance, data protection
Architecture & Delivery	Lead Solution Architect	Technical design, integration, and architecture

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Tier	Role	Primary Focus
	Data Platform Engineer	Data pipelines, ingestion, and catalog integration
	DevOps Engineer	Environment setup, CI/CD, automation
UX & Application Development	UX Designer	User experience and accessibility
	Web Development Engineer	Frontend development and catalog UI
Automation & Operations	ServiceNow Developer	Workflow automation, approvals, lineage tracking

Our delivery team includes the requisite skills and experience required to deliver a seamless, end-to-end implementation. Table 2 describes the key personnel positioned to begin supporting this engagement upon award. Detailed resumes for each key personnel are provided in Appendix A.

5.0. Past Performance

Archetype has successfully delivered robust data integration and automation solutions for clients facing fragmented systems and manual reporting burdens. The example below highlights our work at Vermont Health Connect (VHC) with the State of Vermont and demonstrates our ability to implement secure, scalable pipelines and automated refresh cycles, in alignment with the scope of work outlined in the RFP.

Project Name: EOTSS Enterprise Data Governance Program and Data Intelligence Platform (DIP) Implementation	
Client/Organization	Executive Office of Technology Services and Security (EOTSS), Office of Data Innovation and Analytics (ODIA), Commonwealth of Massachusetts
Contact Name and Title	Josh Merrill, Director of Enterprise Data Analytics
Email Address	Joshua.Merrill@mass.gov
Phone Number	(617) 671-5564
Relevance to the Massachusetts AI Hub Data Commons	
This engagement established the Commonwealth's enterprise governance, metadata management, and interoperability foundation, which directly supports the MassTech AI Hub Data Commons Collaborative (DCC) objectives. Through the **EOTSS Enterprise Data Governance Program and Data Intelligence Platform (DIP)**, Archetype operationalized EOTSS's enterprise governance vision creating standardized frameworks for data stewardship, policy enforcement, lineage tracking, and cross-agency data sharing. The Executive Office of Labor and Workforce Development (LWD) served as the initial secretariat rollout and pilot deployment for the DIP project introducing the first statewide model for metadata driven governance, integrating governance policies and workflow automation across Snowflake, ServiceNow, and Alation to ensure that all Commonwealth data assets are managed consistently, ethically, and in compliance with FAIR principles. The success of this first rollout validated the model	

Use or disclosure of data contained on this page is subject to the restriction on the title page.

now being replicated across all secretariats and serves as the blueprint for the DCC's governance framework and metadata catalog design.

Scope and Objectives

Archetype was selected by EOTSS to establish and operationalize the Enterprise Data Governance Program and the Data Intelligence Platform (DIP) across the Commonwealth. The goal was to move the state from siloed data management to a federated, governed, and interoperable model that supports analytics and AI initiatives statewide.

Key objectives included:

- Developing and institutionalizing a statewide governance framework defining governance groups, roles, and decision-making authority.
- Implementing metadata-driven governance workflows that automate dataset onboarding, access approvals, and lineage tracking.
- Configuring Alation as the enterprise metadata catalog and integrating it with ServiceNow for access request automation and audit logging.
- Establishing EOTSS governance standards for data sharing, privacy, and compliance, applied first at LWD to test repeatability and scalability.
- Deploying executive and leadership governance groups at LWD (Data Governance Executive Group and Data Governance Lead Group) with formal charters, escalation paths, and performance measures.
- The engagement provided the operational governance model that now anchors the Commonwealth's AI governance and DCC metadata structure.

Solution and Technical Approach

Archetype implemented a cloud-native, federated governance architecture leveraging the Commonwealth's existing enterprise technologies.

Archetype applied its **ICON Methodology** (Identify, Calculate, Optimize, Navigate) to guide implementation and agency adoption. At LWD, Archetype delivered the first operational deployment integrating ServiceNow and Alation workflows, mapping data stewardship roles, and training governance teams to ensure long-term sustainability. This secretariat-level proof-of-concept demonstrated the DIP's scalability across the enterprise and established its governance structure as the model for future agencies.

The Data Intelligence Platform (DIP) integrates:

- **Snowflake**: Enterprise data hosting and sharing platform for governed, secure, and scalable data management.
- **AWS**: Infrastructure environment ensuring compliance, performance, and elasticity for large-scale data processing.
- **ServiceNow**: Workflow automation for access controls, stewardship approvals, and audit compliance.
- **Alation**: Metadata and cataloging platform that operationalized FAIR principles, providing data lineage and discoverability across agencies.

Results and Outcomes

- **Operationalized the first statewide governance framework**, deployed at LWD, with active executive and leadership governance groups.
- **Implemented FAIR-aligned metadata cataloging** with Alation, standardizing metadata for over 60 enterprise datasets.

Use or disclosure of data contained on this page is subject to the restriction on the title page.

- **Improved governance compliance and audit readiness**, enabling automated lineage reporting and cross-agency data traceability.
- **Established the Commonwealth's enterprise governance and metadata model**, which now serves as the foundation for the AI Hub Data Commons Collaborative.

Project Name: **Defense Health Agency (DHA) – Biosurveillance Program (B-HIVE/BIP)**	
Client/Organization	Defense Health Agency (DHA)
Contact Name and Title	Neils Olson, Chief Medical Officer of Defense Innovation Unit (DIU)
Email Address	nolson@diu.mil
Phone Number	
Relevance to the Massachusetts AI Hub Data Commons	
The DHA BIP demonstrates how an open-standards-based catalog (data.world) integrated with ServiceNow workflows and advanced AI analytics can deliver real-time, federated intelligence in a mission-critical context. Just as the DHA required a secure, auditable platform for Biosurveillance, MassTech's AI Hub Data Commons requires an open, governed, and explainable foundation that supports community contributions, AI-driven analytics, and transparent decision-making. This past performance proves the viability of deploying such an architecture quickly, securely, and at scale.	
Scope and Objectives	

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Project Scope

The Defense Health Agency (DHA) requires the development and operationalization of an integrated **Biosurveillance Information Platform (BIP)** to enhance early detection, tracing, and mitigation of biological threats. The platform must unify structured and unstructured Biosurveillance data, apply explainable artificial intelligence (AI) and machine learning (ML) models, and deliver transparent, auditable insights to support rapid operational decision-making across the Department of Defense (DoD) and interagency partners.

The BIP initiative will establish a **governed data and analytics ecosystem** capable of:

- Aggregating diverse Biosurveillance data streams (e.g., wastewater, mobility, clinical, open-source intelligence).
- Enabling real-time situational awareness through federated data discovery and contextual analytics.
- Supporting traceable, evidence-based decision-making with reproducible model outputs and provenance tracking.

The project encompasses the design, configuration, and deployment of core system components including the semantic data catalog, metadata governance layer, AI/ML explainability framework, and automated workflow management for data and analytic product lifecycle governance.

Project Objectives

1. **Integrate Fragmented Data Sources**
   Establish secure, automated pipelines to ingest and harmonize structured and unstructured Biosurveillance data (e.g., reports, PDFs, media streams, clinical datasets) into a unified metadata-driven platform.
2. **Implement a Federated Metadata and Knowledge Graph Framework**
   Develop and deploy a semantic metadata layer (using RDF/RDFS) that links assets across data silos, ensuring interoperability and supporting standardized governance and FAIR data principles.
3. **Deliver Explainable AI Capabilities**
   Implement an AI Context Engine that automatically pairs analytic outputs with underlying data and model metadata, enabling full transparency and traceability of assumptions, features, and training data used in decision-making.
4. **Automate Data Governance and Lineage Tracking**
   Integrate workflow automation tools (e.g., ServiceNow Data Fabric) to manage approvals, lineage validation, and compliance verification for all analytic and data assets.
5. **Develop Advanced Biothreat Modeling and Simulation Tools**
   Deploy a suite of AI and ML models (e.g., GCLASS situational awareness engine, SIRf stochastic simulator, Bayesian inference models) to generate real-time epidemiological insights and evaluate potential interventions across scenarios.
6. **Enable Operational Decision Support and Visualization**
   Provide leadership with interactive dashboards and simulation interfaces that summarize model outputs, visualize risk levels, and quantify potential outcomes of countermeasures such as flight restrictions, mask mandates, or containment zones.
7. **Ensure Security, Provenance, and Auditability**
   Embed role-based access control (RBAC), metadata provenance tracking, and audit logging throughout the platform to ensure compliance with DoD data protection and reporting requirements

Solution and Technical Approach

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Working with Cyberhill and KUNGFU.AI, data.world serves as the semantic catalog and governance backbone for the BIP. Key capabilities include:

- Federated Metadata & Knowledge Graphs – data.world integrates structured datasets (e.g., MITRE wastewater, GDELT news, mobility data) with unstructured sources via autonomous ingestion agents. RDF/RDFS graphs describe, link, and govern assets across silos.
- AI Context Engine for Explainability – Every model output is paired with contextual metadata, enabling "show your work" explainability. Analysts can interrogate model assumptions and trace outputs to underlying data.
- Workflow Automation – ServiceNow Workflow Data Fabric automates approvals, lineage tracking, and governance checks, ensuring that all analytic products are transparent, auditable, and compliant.
- Advanced Biothreat Analytics – KUNGFU.AI models and simulations deployed within the BIP include:
  - GCLASS: LLM-based real-time situational awareness engine that scrapes, translates, and classifies global news feeds into structured epidemiological insights.
  - SIRf Flight-Network Model: Stochastic simulator forecasting pathogen spread across global air routes and testing interventions such as flight lockdowns or mask mandates.
  - Bayesian Inference Models: Capable of tracing outbreak origins with over 75% accuracy across the first 100 simulated pathogens

Results and Outcomes

Enhanced Detection: The BIP demonstrated early detection of novel threats, fusing wastewater, mobility, and media feeds into a live operational dashboard

- Decision Support: Countermeasure simulations now provide quantified projections for interventions, enabling leadership to test "what if" scenarios in real time.
- Traceable Intelligence: All analytic outputs are governed in data.world with lineage metadata, provenance, and workflow audit trails, giving decision-makers confidence in the evidence base.
- Scalability: Live deployment on cloud infrastructure has scaled to ingest an expanded set of data sources, with live alerting, algorithmic event detection, and automated analytic reporting

Enhanced Detection: The BIP demonstrated early detection of novel threats, fusing wastewater, mobility, and media feeds into a live operational dashboard.
Decision Support: Countermeasure simulations now provide quantified projections for interventions, enabling leadership to test "what if" scenarios in real time.
Traceable Intelligence: All analytic outputs are governed in data.world with lineage metadata, provenance, and workflow audit trails, giving decision-makers confidence in the evidence base.
Scalability: Live deployment on cloud infrastructure has scaled to ingest an expanded set of data sources, with live alerting, algorithmic event detection, and automated analytic reporting.

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Project Name: Vermont AI Essential Case Summary (ECS) "SnowFlow" and Modern Data Analytics & Reporting (MDAR) Platform Implementation	
Client/Organization	Department of Vermont Health Access (DVHA) – Vermont Health Connect (VHC)
Contact Name and Title	Daniel Fay, DVHA Deputy Director of Operations
Email Address	daniel.fay@vermont.gov
Phone Number	(802) 657-7523

Relevance to the Massachusetts AI Hub Data Commons

Archetype's Vermont work delivers the same foundations the DCC requires: a secure, governed data platform and a controlled path to applied AI. We implemented the MDAR analytics backbone on Snowflake/AWS with role-based access, encryption, monitoring, and documented ELT models that support metadata-driven reporting and FAIR-oriented discovery. On top of that governed warehouse, we built Essential Case Summary (ECS), a template-driven, auditable LLM workflow that produces standardized summaries from trusted tables (not raw prompts), reinforcing controls and data stewardship. Together, these outcomes align directly with the DCC's priorities for secure hosting and access, catalog/metadata practices, and rapid, governed AI prototyping suitable for multi-stakeholder use across state programs.

Scope and Objectives

Archetype was engaged to design, build, and continuously enhance Vermont's enterprise reporting and analytics platform supporting health and human services operations.
Key objectives included:

- Stand up a secure, cloud-based analytics environment on AWS/Snowflake for eligibility, enrollment, and reporting operations.

Migrate and modernize reporting from legacy systems to curated ELT models with documented lineage and refresh schedules. Establish governance and security controls aligned to NIST 800-53 and MARS-E, and support required federal reporting to the IRS and CMS).Deliver role-based dashboards and audit-ready reports to improve operational visibility, accountability, and decision speed. Implement Essential Case Summary (ECS)/SnowFlow, a template-driven & governed LLM workflow, to standardize case documentation and reduce review/handle time.

Solution and Technical Approach

Archetype implemented a secure, AI-driven data ecosystem combining Snowflake, AWS, and Tableau with the Modern Data Analytics & Reporting (MDAR) platform. The system's design reflects an architecture similar to the DCC's planned model emphasizing governance, transparency, and ethical AI adoption. At the center of the platform is SnowFlow (ECS), an LLM-powered module that automates case summarization by transforming transcripts, agent notes, and structured data into standardized one-page tailored summaries, improving call center operations. The MDAR platform harmonizes data from eligibility, payment, and case management sources through automated ELT pipelines refreshed twice daily, ensuring operational accuracy and real-time analytics.

The environment uses a structured metadata catalog with documented lineage, refresh schedules, and role-based access controls. Access requests and approvals are handled through standard workflow procedures with auditable change control. Deployed on AWS in an operating posture aligned to MARS-E, the platform applies end-to-end encryption, continuous monitoring, and automated audit logging. This approach meets the DCC's expectations for secure, explainable, and metadata-driven AI use across multi-agency programs.

Primary Technologies: Snowflake, AWS, Tableau, Archetype SnowFlow AI Module (LLM/NLP), MDAR Accelerator

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Results and Outcomes
• ECS/SnowFlow AI solution reduced call center hold time by 80-90%, from 5-10 minutes to ~1 minute per case. • Improved reporting performance by 99%, reducing runtime from hours to seconds. • Potential annual cost savings exceeding $500K through process automation. • Achieved >99.99% uptime with continuous availability and zero audit exceptions. • Delivered a repeatable, scalable architecture now serving as a reference model for AI adoption in other state systems. The MDAR platform implementation, paired with the Essential Case Summary (ECS/SnowFlow) capability, demonstrates Archetype's ability to design, operationalize, and sustain a secure, AI-enabled data environment that aligns with the MassTech AI Hub Data Commons goals. The work shows proven strengths in metadata governance, role-based access, auditability, and ethical, template-driven AI over trusted warehouse data providing a tested pattern the DCC can adapt for statewide use.

6.0. Closing Statement

Archetype, together with SMX and data.world, is committed to delivering a secure, scalable, and operationally ready Data Commons for the Massachusetts AI Hub. As the lead integrator, Archetype brings public sector program leadership and hands on delivery experience from projects like MDAR with ECS and SnowFlow in Vermont to establish governed data hosting on AWS and Snowflake, a robust metadata foundation, role-based access, and a controlled path for applied AI. SMX strengthens the team with enterprise cloud engineering, security hardening, monitoring, and day two operations at scale. data.world provides a modern, standards aligned catalog and knowledge graph for search, lineage, stewardship, and discoverability across agencies.

Our joint approach is practical, and outcomes driven, with clear milestones, traceable deliverables, and performance measures that matter, including search responsiveness, refresh service levels, adoption, and time to insight. We will partner closely with MassTech and stakeholders to finalize a fit for purpose roadmap, manage risk, and transfer capability so the Commonwealth can sustain and extend the platform. With complementary strengths, named specialists, and ready to execute plans for discovery, credits, security, and training, the team of Archetype, SMX, and data.world is prepared to start immediately and deliver a Commons agencies and researchers trust and that the AI Hub can confidently scale statewide.

7.0. APPENDIX A: Key Personnel Resumes

Augusto Rosa, Lead Solution Architect

Name	Augusto Rosa		
Title	Lead Solution Architect	**Years of Relevant Experience**	10+
Degree/Education	• Bachelor of Science in Computer Science, Universidade Agostinho Neto		
Certifications	• AWS Certified Solutions Architect • CISSP – Certified Information Systems Security Professional • Recognized Snowflake Data Superhero		
Professional Summary			

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Name	Augusto Rosa

Augusto Rosa is the Head of Data, Cloud, and Security Architecture at Archetype Consulting. With over a decade of experience in enterprise cloud strategy, data architecture, and platform engineering, Augusto specializes in helping organizations modernize their data platforms through secure, scalable cloud solutions. A recognized Snowflake Data Superhero and experienced public speaker, he brings deep technical expertise in AWS, Snowflake, Redshift, and infrastructure-as-code. Augusto excels at translating business needs into architectural blueprints, leading both technical strategy and implementation for high-impact data and analytics initiatives.

Relevant Experience

Archetype Consulting
Head of Data, Cloud, and Security Architecture | Feb 2025 - Present

- Leads enterprise-scale data platform, security, and cloud architecture strategy
- Co-architects scalable cloud infrastructure and governance frameworks
- Authored and published a Snowflake Architecture foundational checklist (May 2025)
- Developed internal thought leadership content on Snowflake features and data lakes

Infostrux Solutions
Solutions Architect and Data Engineering Lead | Feb 2021 – Feb 2025

- Built the company's data engineering practice
- Recognized twice as a **Snowflake Data Superhero** (2024 & 2025)
- Designed and executed Snowflake and Redshift migrations, modernization, and alert-driven analytics

Slalom Consulting
Cloud and Platform Engineer and Architect | 2017-2024

- Early career spent architecting cloud-native systems and automation for diverse clients
- Specialized in AWS, infrastructure-as-code, CI/CD pipelines, and DevOps enablement
- Led multi-cloud builds and SRE-aligned platform engineering initiatives

Shomi
Director of Infrastructure and Operations | 2014-2016

- Directed Infrastructure, DevOps, Digital Operations, Business IT, and Video Operations teams for a national streaming platform
- Oversaw end-to-end platform delivery, including systems uptime, streaming quality, and customer-facing performance
- Developed and executed infrastructure strategies aligned to scalability, security, and cost efficiency
- Drove cross-functional initiatives across IT, security, and product operations to support high-availability content delivery
- Implemented cybersecurity and information security measures to support platform resilience
- Built and maintained strong internal team relationships and vendor partnerships to support growth and innovation
- Advised on emerging technology adoption and platform modernization opportunities
- Other Relevant Experience:
- Led a large-scale enterprise data warehousing and executive KPI dashboard implementation for a Software-as-a-Service (Saas) provider
- Solutions Architect on a strategic project with a SaaS firm to move from a legacy operational reporting platform to cloud data warehouse and advanced visualizations

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Canon Patillo, AI Governance Architect

Name	Canon Patillo		
Title	AI Governance Architect	**Years of Relevant Experience**	12+
Degree/ Education	• P.E., Digital Transformation, Massachusetts Institute of Technology, Cambridge, MA • MBA, University of Maryland University College, Adelphi, MD • MS, Data Science, University of Maryland University College, Adelphi, MD • BS, International Relations, United States Naval Academy, Annapolis, MD		
Certifications	• Project Management Professional (#1679446), Project Management Institute		

Professional Summary

Canon Pattillo is the Senior Director, AI Strategic Transformation and Implementation at Archetype Consulting. With over twelve years of experience spanning data science, advanced analytics, and organizational change, he leads the design and execution of AI governance programs that operationalize fairness, transparency, and security across complex data ecosystems. A U.S. Navy veteran and seasoned AI practitioner, Canon blends hands-on machine learning expertise with strategic program design to help public and private organizations adopt responsible, high-value AI.

At Archetype, Canon architects and implements enterprise AI and data governance frameworks for public and private organizations. His work integrates Zero Trust principles, model risk management, and transparent decision workflows to ensure accountability from data ingestion through model deployment. Canon's approach unites technical rigor with pragmatic change management aligning technology, people, and policy to build sustainable, explainable, and trusted AI capabilities.

Relevant Experience

Archetype Consulting
Senior Director, AI Strategic Transformation and Implementation | 2024–Present

- Built and implemented the enterprise AI strategy and operating model for a private organization, defining governance structures, stewardship roles, and adoption pathways that embedded AI ethics, security, and accountability into business operations.
- Led the design and execution of AI governance frameworks that operationalize fairness, transparency, and Zero Trust principles across data ingestion, model development, and deployment lifecycles.
- Directed cross-functional AI programs integrating data strategy, responsible AI policy, and automation accelerating time-to-value while ensuring auditability and compliance with emerging governance standards.
- Championed cultural and operational change management including training stakeholders, defining decision rights, and scaling responsible AI practices across business and technical domains.

Porter
Vice President, Informatics and Data Science | 2021–2024
Vice President, Supply Chain | 2021–2024

- Founded and led the Informatics and Data Science vertical, establishing Porter's data governance, analytics, and machine learning strategy from inception through enterprise adoption.
- Designed and deployed production-grade ML pipelines, recommendation engines, and LLM-enabled prototypes using Python, Databricks, AWS SageMaker, and LangChain to improve prediction accuracy and operational transparency.

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Name	Canon Patillo

- Developed end-to-end data lifecycle governance policies integrating PHI handling, data access control, and zero-trust security protocols across multiple internal and client platforms.
- Directed all stages of model lifecycle management (design, evaluation, deployment, and monitoring) ensuring explainability and auditability in compliance-sensitive environments.
- Built and operationalized Porter's R&D Lab, enabling rapid experimentation and model validation through reproducible, policy-controlled data environments.
- Led data partnerships and vendor integrations, curating and harmonizing multi-industry datasets to support ethical, compliant AI development.

Procter & Gamble
Analytics and Insights IT Director | 2020–2021

- Led the Online Pickup and Delivery (OPD) Analytics and Insights portfolio, driving data quality and predictive analytics initiatives that improved product availability from 77% to 90%.
- Directed agile development of the Share-of-Shipments tool and Rapid Response platform, incorporating fairness-aware alerting and data transparency mechanisms for global supply-chain visibility.
- Delivered cross-functional data governance enhancements, integrating data provenance and metadata tagging across the enterprise DataHub environment.
- Established standards for model validation and data accuracy within the P&G-Walmart analytics ecosystem, strengthening trust in operational dashboards and AI-driven insights.
- Enabled the first global rollout of the Walmart International DataHub implementing secure, governed access to omni-channel datasets for 90% of P&G product lines.
- Championed data-driven decision-making and automated monitoring workflows that reduced manual report rebuilds by 65%.

InflectionLearning
Founder and Learning Architect | 2019–2020

- Founded and directed a data-driven ed-tech startup focused on adaptive learning models, data ethics, and equitable student outcomes.
- Designed and implemented AI-based assessment engines to deliver individualized instruction while maintaining student privacy and data-security compliance.
- Established governance practices for educational data use, including consent management, metadata curation, and transparent performance reporting.
- Oversaw data pipeline architecture and analytics integrations used by pilot schools across three states, improving engagement metrics and learning outcomes.
- Built a ten-person cross-functional team emphasizing data literacy, innovation, and ethical AI deployment in classroom environments.
- Managed seed funding and stakeholder communication, ensuring responsible scaling through evidence-based decision frameworks.

Pulse8
Senior Director, BI and Analytics | 2016–2019
Director, Business Intelligence | 2014–2016

- Established and led Pulse8's enterprise data governance and analytics organization, reducing ETL errors by 82% and improving transparency across health-plan data pipelines.
- Directed creation of predictive analytics products supporting risk-adjustment and quality measurement for major healthcare payers.
- Implemented governance frameworks for model validation, data lineage tracking, and accuracy auditing: precursors to formal AI governance protocols.

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Name	Canon Patillo

- Reduced monthly deliverable turnaround time by 31% through automation, metadata standardization, and workflow optimization.
- Delivered a $5M change-management initiative to modernize data integration and visualization across production systems, embedding continuous quality oversight.
- Mentored data science teams in ethical modeling practices, emphasizing explainability and fairness in patient-level analytics.

Mark Cardarelli, Project Manager, Sr.

Name	Mark Cardarelli		
Title	Project Manager, Sr.	**Years of Relevant Experience**	18+
Degree/Education	• Boston College, College of Arts & Sciences – Chestnut Hill, MA • Bachelor of Arts in Political Science, Minor in Philosophy, Class of 2006		
Certifications	• Oracle Certified Implementation Specialist – Hyperion Planning • Oracle Certified in Professional Services • Oracle Certified in Planning & Budgeting Cloud Services		

Professional Summary

Results-driven Technical Project Manager and Analytics Leader with 18+ years of experience guiding complex cloud, data, and business-intelligence programs from strategy through delivery. Proven success leading multi-vendor teams across the Commonwealth of Massachusetts to modernize analytics environments using Snowflake, AWS, and Alation under the EOTSS Enterprise Analytics Program. Recognized for enabling cross-secretariat collaboration, advancing data governance maturity, and creating reusable, governed data products that support economic development, workforce, and housing policy decisions statewide.

Relevant Experience

Archetype Consulting
Senior Director, Analytics Practice | March 2021 - Present

- Lead project management and delivery for the EOED Data Warehouse Phase 3, a cornerstone of the Commonwealth's Enterprise Analytics Program focused on integrating economic, workforce, and transportation data within a federated Snowflake architecture.
- Coordinate multi-agency teams (EOTSS, EOED, SMX, Archetype) to execute an agile, backlog-driven modernization roadmap, ensuring alignment with ODIA's (Data Innovation & Analytics Office) data governance and Snowflake standards.
- Manage sprints, work-planning, and deliverables in collaboration with the Commonwealth's project management office; oversee risk, issue, and performance tracking using Smartsheet and JIRA.
- Improved stakeholder satisfaction and supporting Commonwealth AI readiness and data commons initiatives.

Director, Analytics Insights | March 2013 - March 2021

- Directed delivery of large-scale data warehousing and business intelligence initiatives for enterprise clients.
- Managed client relationships, requirements gathering, and technical design processes to align data strategies with organizational goals.
- Implemented cloud-based ELT pipelines integrating disparate data sources to enable real-time analytics and self-service reporting.
- Mentored consulting teams and developed frameworks for repeatable project delivery.

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Name	Mark Cardarelli

- Achieved multiple Oracle certifications (Hyperion Planning, Professional Services, Planning & Budgeting Cloud Services).

Manager, Analytics Insights | January 2012 - March 2013

- Oversaw small teams delivering analytics and reporting solutions using Oracle and emerging cloud tools.
- Coordinated internal resources and vendors to ensure projects were delivered on time and within budget.

Senior Consultant | September 2007 – January 2012

- Supported full lifecycle design and implementation of data-driven business solutions for mid- to large-scale clients.

Jim Waldrop, Advisory Lead

Name	Jim Waldrop		
Title	Advisory Lead	**Years of Relevant Experience**	20+
Degree/Education	• Master of Strategy Studies, U.S. Army Senior Service College • Master of Science, University of Texas • Bachelor of Science, Stephen F. Austin State University		
Certifications	• AWS Certified Cloud Practitioner – Issued Oct 20204, Exp. 2027		

Professional Summary

A results-driven technology leader with over 20 years of experience in IT governance, cloud compliance, and risk management. Expert in developing and implementing comprehensive cloud governance frameworks within highly regulated industries, particularly financial services. Proven ability to lead cross-functional teams and collaborate with senior leadership to optimize cloud operations, ensure regulatory compliance, and mitigate risk. Adept at working with multi-cloud environments, including AWS and Azure, while driving operational efficiency and data security through governance best practices. Strong analytical, problem-solving, and communication skills with a passion for continuous improvement.

Relevant Experience

Archetype Consulting
Sr. Director of Advisory | October 2024 – Present
Head of Data Governance for Boston Children's Hospital (BCH)

- Established and led the hospital's **Data Governance Office (DGO)**, managing a cross-functional team of data stewards, analysts, and technical leads responsible for data quality, stewardship, and policy implementation across clinical and operational domains.
- Designed and operationalized the **enterprise Data Governance Framework**, aligning governance processes with regulatory standards (HIPAA, SOC 2, GDPR) and BCH's strategic data and analytics objectives.
- Directed the **implementation of the enterprise data catalog (Alation)**, defining metadata standards, stewardship workflows, and taxonomy for domains spanning Epic, Mosaic, financial, and research systems.
- Authored and implemented **governance policies, stewardship playbooks, and data classification models** to standardize ownership, lineage, and access management across the enterprise.
- Partnered with senior leadership to develop and enforce **data quality metrics, lineage transparency, and metric definition standards**, reducing report redundancy and improving analytic trust across departments.

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Name	Jim Waldrop

- Launched and chaired **Data Governance Council and Working Groups**, fostering collaboration between IT, compliance, and business units to ensure unified policy execution and governance maturity.
- Developed and tracked **key performance indicators (KPIs)** for data governance adoption, catalog utilization, and stewardship engagement, achieving measurable improvements in metadata completeness and user participation.
- Led **change management and training programs**, enhancing hospital-wide literacy on data stewardship, governance processes, and responsible data use.
- Provided strategic alignment between data governance, **Epic analytics modernization**, and enterprise initiatives such as **Project Unify and Mosaic optimization** to enable scalable, FAIR-aligned data practices.

Commercial Director A.T. Kearney | July 2022 – September 2024

- Developed and implemented robust cloud governance strategies for Global Financial Services Corporations that aligned IT and business objectives, ensuring alignment with FDIC, GDPR, FIPS compliance.
- Collaborated with senior leadership to deliver compliance updates, risk assessments, and governance enhancements, resulting in a 20% improvement in operational efficiency.
- Led the implementation of governance policies in multi-cloud environments, achieving full compliance with regulatory requirements and minimizing risk.
- Developed and maintained IT governance policies ensuring compliance with industry regulations and internal standards.
- Led the risk management initiatives, identifying and mitigating risks associated with IT operations and third-party vendors for various organizations.

Slalom Consulting
Client Service Partner, Account Lead | January 2020 – July 2022

- Acted as a key advisor to executive leadership, providing insights on cloud governance matters and ensuring policies supported business objectives while maintaining security.
- Led a team of 10 cloud governance specialists, providing mentorship, training, and guidance, leading to a 30% increase in team productivity and compliance adherence.
- Developed cloud cost management initiatives that led to a 10% reduction in overall cloud expenditures through optimized governance practices.
- Launched data migration products, migrating 1PB of protected data to AWS.
- Created process/owner Roadmaps, process activities, and controls for a social media client with over 3.6 billion monthly users, ensuring compliance with privacy standards.

Booz Allen Hamilton
Sr. Director, IT Governance | June 2005 – December 2019

- Implemented automation tools and processes to enhance governance, compliance monitoring, and risk management, increasing efficiency by 35%.
- Stayed abreast of industry trends, adapting governance strategies to meet evolving regulatory requirements and cloud technologies.
- Promoted cloud governance thought leadership within the organization, enhancing the company's reputation in data compliance and security.
- Managed a global team of 33 IT professionals, growing the team by 18% over three years.
- Earned the MVP Award, the companies highest award, exemplifying Booz Allen's core values of passionate service, unflinching courage, ferocious integrity, and collective ingenuity.

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Glen Trudeau, DevOps Platform Engineer

Name	Glen Trudeau		
Title	DevOps Platform Engineer	**Years of Relevant Experience**	10+
Degree/Education	• Bachelor of Applied Arts and Sciences (BAAS) in Organizational Leadership degree at Texas A&M University-Commerce		
Certifications	• AWS Certified Security – Specialty • AWS Certified Developer - Associate • AWS Certified Cloud Practitioner • Snowflake Associate Architect • Microsoft Certified: Azure Fundamentals • AWS Partner: Cloud Economics Accreditation • AWS Partner: Accreditation (Technical) • AWS Partner: Accreditation (Business)		

Professional Summary

Glen is an experienced cloud infrastructure engineer, skilled in designing, delivering, and supporting stable complex cloud-based infrastructure projects based in Amazon Web Services, Azure, Google Cloud Platform, Hybrid cloud etc. In this role Glen displays a strong understanding of release management, particularly CI/CD tooling and systems, and has demonstrated his commitment to integrate security best practices and cost savings into core functionality. Glen has experience with architecting data flows from a number of sources, including Oracle, Snowflake, SQL Server, and Redshift, as well as reporting tools like Birst, Tableau, and Power BI.

Relevant Experience

Archetype Consulting
Manager, Cloud Engineering | October 2021 - Present

- Migration of servers and systems to cloud infrastructure resulting in increased stability and lower costs.
- Cloud containerized stack build focused on scraping market data used for market analysis.
- Regional infrastructure failover planning/design/testing
- Migration of existing cloud stacks to Terraform management for DR and replication
- CDN Management
- Implementation of full stack monitoring and real time error handling/alerts
- Stack automation and data load design
- Full stack security monitoring and alerting with tools like Data Dog, Cloud Watch, Cloud Trail, Config, Security Hub, Guard Duty, Orca, Shield, Inspector, etc.

DevOps

- Complete automation of deployment process across technology and front/backend teams of cloud infrastructure/web design/content creation
- Stack promotion across Dev/Test/Prod
- Implementation of zero downtime deployments through a blue/green structure

Technology/Infrastructure Leadership

- Developed a complete process for infrastructure ownership that includes monthly security audits, cost management, issue response, and failover. This along with constant technical support at all hours has kept the infrastructure of multiple projects running smoothly for both project leads and clients.

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Candice Sutterfield, Data Platform & AI Lead

Name	Candice Sutterfield		
Title	Data Platform & AI Lead	**Years of Relevant Experience**	20+
Degree/Education	• BBA, Management Information Systems at Texas A&M University • MS, Master of Science Business Analysis at East Texas A&M University		
Certifications	• Snowflake SnowPro Core • Dataiku ML Practitioner • Databricks Generative AI Fundamentals • OneStream Sensible Machine Learning		

Professional Summary

Candice Sutterfield is a Data Scientist with over 20 years of experience in business process improvement, data strategy, and analytics. With her deep technical expertise and strategic vision, Candice is passionate about the intersection of AI, people, and processes helping organizations unlock the full potential of their data to drive innovation and informed decision-making. Her AI and data expertise includes generative LLM solutions, predictive modeling, time series forecasting, data architecture, data engineering, and business intelligence. Candice has delivered analytic solutions across industries, including Non-Profit, Insurance, Financial Services, and Banking. She is well-versed in leading technologies such as Snowflake, Python, SQL, Dataiku, Databricks, DataRobot, Tableau, and ServiceNow.

Relevant Experience

Archetype Consulting
Data Analytics and AI Leadership & Delivery | August 2022 - Present

- Organize and deliver zero-trust security configuration for databases.
- Build data collaboration solutions to share information across organizations securely and efficiently.
- Develop statistical models and comprehensive visualizations for enhanced customer analytics and insights
- Deliver Gen AI solution for government health care organization that summarized relevant case details for call center reps.
- Solution design of enterprise-wide Data Warehouses for analytics performed by Finance, Marketing, and Business Decisions support functions
- Data integration and visualization tool evaluation and implementation
- Data Governance, including source definition, data standardization, data model design, report rationalization and validation.

Customer Engagement

- Develop analytic solutions that ensure customer and organizational goals are met
- Establish and lead analytic working groups that represent desperate perspectives
- Champion user adoption of solutions; develop and deliver change management and communication plans in addition to training assets (webinars, user guides, etc.)
- Prosci Change Management certification which focuses on managing the people side of change to deliver results-focused solutions

Goktug Erkan, Security Architect

Name	Goktug Erkan		
Proposed Position	Security Architect	**Years of Relevant Experience**	12+

Use or disclosure of data contained on this page is subject to the restriction on the title page.

IRS MBI Status	No current IRS MBI but can obtain an MBI clearance. Currently holds an active HHS NIH NCI/NHLBI, HHS SAMHSA, PBGC, and DOL Public Trust clearances.
Degree/Education	Master of Science in Cybersecurity – Marymount University (VA) Bachelor of Science in Computer Science and IT – Marymount University (VA)
Certifications	CISA, CEH, CISSP, and Security+

Professional Summary

Mr. Erkan is an experienced cybersecurity professional with over eight (8) years of expertise in the field and an additional four (4) years in secure software development lifecycle (SSDLC). He has led and managed various cybersecurity projects for Federal, state, and commercial customers, providing RMF, ATO, security, compliance, cloud, and SSDLC consultancy and implementation services. Mr. Erkan has extensive experience in various types of cybersecurity and risk assessments, vulnerability assessments, penetration testing, design reviews, and evaluations, as well as pre- and post-assessment support. He has maintained two (2) Public Trust clearances and has worked with various regulatory agencies, including NIH, NHLBI, NCI, SAMSHA, CMS, FCC, USAC, ED, FSA, PBGC, DOL, and state governments to deliver successful project deliverables and obtaining ATO grants. Holds active HHS, DOL, PBGC Public Trust clearances.

Relevant Experience

Archetype Consulting
Director, Compliance & Security | October 2021 - Present

- Successfully managed all pre- and post-assessment cybersecurity efforts for FISMA-based Authorization to Operate (ATO) and MARS-E-based Authorization to Connect (ATC), resulting in the successful granting of ATOs by NIH, NCI (OneStream), NHLBI (OneStream), and SAMHSA (OneStream), and ATC by CMS. He provides cybersecurity, RMF, and compliance consultancy or implementation services to customers seeking ATOs, cloud migrations, assessments, or new cybersecurity products such as cloud-native tools and concepts like Security Data Lake. He oversees the compliance and security of company delivery and solutions to ensure continuous compliance with various regulations. Serves as ATO Specialist lead for PBGC OS PPBE project as of Sep 2024 and leading ATO for Dept of Labor as of April 2024.

DXC Technology
Senior GRC Engineer; RMF SME & TPM | June 2019 – Oct 2021

- Mr. Erkan conducted security and risk assessments to determine compliance with various regulations, standards, and frameworks such as FISMA, MARS-E, HIPAA, FedRAMP, PCI, NIST, SOC2, and ISO. He developed security roadmaps and ensured compliance with risk management standards, managing relationships with senior management, customers, and partners at all levels of the organization. Mr. Erkan provided subject matter expertise for RMF and ATO projects based on FISMA, FedRAMP, or other NIST 800-53 compliance standards and regulations. He supported Federal ISSOs and participated in cloud security readiness assessments, system security evaluations, internal and external penetration tests, vulnerability scans, security control assessments, audits, and reviews to determine security posture and vulnerabilities. Lastly, Mr. Erkan led various cybersecurity projects: cloud security assessments, security control assessments, vulnerability assessments, and ATO projects for the CMS and state governments.

Nu-Pulse Technologies
Senior Security Engineer | October 2018 – May 2019

- Mr. Erkan provided expertise in obtaining ATOs for various Federal and state government and commercial projects through the application of the RMF. He successfully delivered FISMA or NIST-based re-authorization or new authorization projects on time and on budget, obtaining ATOs from FCC, USAC, and state governments. Mr. Erkan assisted customers with security program development, conducting security assessments, risk assessments, and POA&M to ensure compliance with regulatory frameworks.

Fila Group
Cyber Security Specialist | November 2017 – Oct 2018

- Mr. Erkan provided cybersecurity support to the Department of Education (ED) and FSA's EA program to improve IT investment decision-making and align strategic objectives with business needs and IT resources.

Use or disclosure of data contained on this page is subject to the restriction on the title page.

APPENDIX B: RFP Requirements mapped to Epics based on percent completion

Name	#	Epic 1 (Data Catalog)	Epic 2 (Credits & Upload)	Epic 3 (Synthetic Data)
Webpage	5.1	100%		
Dataset Directory	5.2	100%		
Dataset Detail Pages	5.3	100%		
Public Data Sets	5.4	100%		
Synthetic Data Tools Section	5.5			100%
Massachusetts Public Datasets	5.6	100%		
Community Dataset Submissions	5.7	30% (feedback)	70% (submissions)	
Admin Dashboard	5.8	50% (dataset mgmt)	50% (submission mgmt)	
Additional Requirements	5.9	50%	25%	25%
Data Hosting and Access	6.1	100%		
User Experience Enhancements	6.2	100%		
Centralized Ledger and Tracking Infrastructure	7.1		70%	30% (synth usage)
Earning Commons Credits	7.2		90%	10% (synthetic contributions)
Spending Commons Credits	7.3		90%	10% (synth usage)
Incentivizing Efficient Usage	7.4		50%	50%
Governance and Transparency	7.5		100%	
User Experience & Access Control	7.6		100%	

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Name	#	Epic 1 (Data Catalog)	Epic 2 (Credits & Upload)	Epic 3 (Synthetic Data)
Technical Expectations	7.7		70%	30%
Capability Definition and Benefits	8.1	100%		
Metadata Management	8.2	100%		
Dataset Discovery and Search	8.3	100%		
Access Control and Governance	8.4	100%		
Dataset Publishing and Versioning	8.5	100%		
Interoperability and Integration	8.6	100%		
Data Usage and Metrics	8.7	100%		
Collaboration and Community Features	8.8	70%	30%	
Data Previews and Exploration	8.9	100%		
Example Data Management and Catalog Capabilities for Consideration	8.10	100%		
Criteria for Inclusion	9.1	100%		
Suggested Initial Datasets	9.2	100%		
Capability Definition and Benefits	10.1			100%
Synthetic Data Generation Requirements	10.2			100%
Synthetic Data Generation Tools Samples	10.3			100%
Purpose	11.1	100%		

Use or disclosure of data contained on this page is subject to the restriction on the title page.

Name	#	Epic 1 (Data Catalog)	Epic 2 (Credits & Upload)	Epic 3 (Synthetic Data)
Governance Requirements	12.0	50%	25%	25%
Core Capabilities	12.1	50%	25%	25%
Operational Requirements	12.2	50%	25%	25%
Governance and Legal Compliance	12.3	50%	25%	25%
Technical Integration	12.4	50%	25%	25%
Monitoring and Reporting	12.5	50%	25%	25%
Education and Capacity Building	12.6	40%	30%	30%
Potential AI Bias and Fairness Open-Source Tools for Consideration	12.7	100%		
Access Controls	13.1	60%	20%	20%
Data Protection	13.2	60%	20%	20%
Guiding Principles	14.1	100%		

Note: in some cases, there will be advisory work that is partially complete in an earlier epic. In those cases, we will fill that requirement entirely for the deliverables in that epic, but there will be future deliverables where we will need to perform additional work. For example, for the AI Fairness & Bias Capability requirement, we will complete this task fully for the initial data catalog but will need to adjust as we add in the credit system and synthetic data capabilities.

Use or disclosure of data contained on this page is subject to the restriction on the title page.

APPENDIX C: Archetype's W-9

Form **W-9**
(Rev. March 2024)
Department of the Treasury
Internal Revenue Service

Request for Taxpayer Identification Number and Certification

Go to *www.irs.gov/FormW9* for instructions and the latest information.

Give form to the requester. Do not send to the IRS.

Before you begin. For guidance related to the purpose of Form W-9, see *Purpose of Form*, below.

Print or type. See *Specific Instructions* on page 3.

1 Name of entity/individual. An entry is required. (For a sole proprietor or disregarded entity, enter the owner's name on line 1, and enter the business/disregarded entity's name on line 2.)

Archetype Consulting, Inc.

2 Business name/disregarded entity name, if different from above.

3a Check the appropriate box for federal tax classification of the entity/individual whose name is entered on line 1. Check only **one** of the following seven boxes.

☐ Individual/sole proprietor ☐ C corporation ☑ S corporation ☐ Partnership ☐ Trust/estate

☐ LLC. Enter the tax classification (C = C corporation, S = S corporation, P = Partnership)

Note: Check the "LLC" box above and, in the entry space, enter the appropriate code (C, S, or P) for the tax classification of the LLC, unless it is a disregarded entity. A disregarded entity should instead check the appropriate box for the tax classification of its owner.

☐ Other (see instructions)

4 Exemptions (codes apply only to certain entities, not individuals; see instructions on page 3):

Exempt payee code (if any)

Exemption from Foreign Account Tax Compliance Act (FATCA) reporting code (if any)

(Applies to accounts maintained outside the United States.)

3b If on line 3a you checked "Partnership" or "Trust/estate," or checked "LLC" and entered "P" as its tax classification, and you are providing this form to a partnership, trust, or estate in which you have an ownership interest, check this box if you have any foreign partners, owners, or beneficiaries. See instructions ☐

5 Address (number, street, and apt. or suite no.). See instructions.

180 Canal St., Ste 600

Requester's name and address (optional)

6 City, state, and ZIP code

Boston, MA 02114

7 List account number(s) here (optional)

Part I Taxpayer Identification Number (TIN)

Enter your TIN in the appropriate box. The TIN provided must match the name given on line 1 to avoid backup withholding. For individuals, this is generally your social security number (SSN). However, for a resident alien, sole proprietor, or disregarded entity, see the instructions for Part I, later. For other entities, it is your employer identification number (EIN). If you do not have a number, see *How to get a TIN*, later.

Note: If the account is in more than one name, see the instructions for line 1. See also *What Name and Number To Give the Requester* for guidelines on whose number to enter.

Social security number

			–			–				

or

Employer identification number

4	5	–	0	5	4	4	1	6	1

Part II Certification

Under penalties of perjury, I certify that:

1. The number shown on this form is my correct taxpayer identification number (or I am waiting for a number to be issued to me); and
2. I am not subject to backup withholding because (a) I am exempt from backup withholding, or (b) I have not been notified by the Internal Revenue Service (IRS) that I am subject to backup withholding as a result of a failure to report all interest or dividends, or (c) the IRS has notified me that I am no longer subject to backup withholding; and
3. I am a U.S. citizen or other U.S. person (defined below); and
4. The FATCA code(s) entered on this form (if any) indicating that I am exempt from FATCA reporting is correct.

Certification instructions. You must cross out item 2 above if you have been notified by the IRS that you are currently subject to backup withholding because you have failed to report all interest and dividends on your tax return. For real estate transactions, item 2 does not apply. For mortgage interest paid, acquisition or abandonment of secured property, cancellation of debt, contributions to an individual retirement arrangement (IRA), and, generally, payments other than interest and dividends, you are not required to sign the certification, but you must provide your correct TIN. See the instructions for Part II, later.

Sign Here Signature of U.S. person — Docusigned by: [signature] — Date March 10, 2025

General Instructions

Section references are to the Internal Revenue Code unless otherwise noted.

Future developments. For the latest information about developments related to Form W-9 and its instructions, such as legislation enacted after they were published, go to *www.irs.gov/FormW9*.

What's New

Line 3a has been modified to clarify how a disregarded entity completes this line. An LLC that is a disregarded entity should check the appropriate box for the tax classification of its owner. Otherwise, it should check the "LLC" box and enter its appropriate tax classification.

New line 3b has been added to this form. A flow-through entity is required to complete this line to indicate that it has direct or indirect foreign partners, owners, or beneficiaries when it provides the Form W-9 to another flow-through entity in which it has an ownership interest. This change is intended to provide a flow-through entity with information regarding the status of its indirect foreign partners, owners, or beneficiaries, so that it can satisfy any applicable reporting requirements. For example, a partnership that has any indirect foreign partners may be required to complete Schedules K-2 and K-3. See the Partnership Instructions for Schedules K-2 and K-3 (Form 1065).

Purpose of Form

An individual or entity (Form W-9 requester) who is required to file an information return with the IRS is giving you this form because they

Cat. No. 10231X — Form **W-9** (Rev. 3-2024)

Use or disclosure of data contained on this page is subject to the restriction on the title page.

APPENDIX D: Attachment B - Authorized Respondent's Signature and Acceptance Form

UNCLASSIFIED//PROPRIETARY
Archetype

9.0 APPENDIX D: Attachment B - Authorized Respondent's Signature and Acceptance Form

Attachment B
Massachusetts Technology Collaborative
Authorized Respondent's Signature and Acceptance Form

The undersigned is a duly authorized representative of the Respondent listed below. The Respondent has read and understands the RFP requirements. The Respondent acknowledges that all of the terms and conditions of the RFP are mandatory, and that Respondent's response is compliant with such requirements.

The Respondent understands that, if selected by the Mass Tech Collaborative, the Respondent and the Mass Tech Collaborative will execute an Agreement specifying the mutual requirements of participation.

Respondent agrees that the entire bid response will remain valid for sixty (60) days from receipt by the Mass Tech Collaborative.

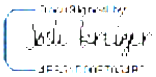
I certify that Respondent is in compliance with all corporate filing requirements and State tax laws.

I certify that the statements made in this response to the RFP, including all attachments and exhibits, are true and correct to the best of my knowledge.

I acknowledge and understand the procedures for handling confidential materials submitted to the Mass Tech Collaborative as set forth in Section 3.1(d); I agree to be bound by those procedures and agree that the Mass Tech Collaborative shall not be liable under any circumstances for the disclosure of any materials submitted to the Mass Tech Collaborative pursuant to this RFP or upon the Respondent's selection.

Respondent: Josh Kreiger
(Printed Name of Respondent)

By:
(Signature of Authorized Representative)



Name: Josh Kreiger
Title: SVP Public Sector & Operations
Date: October 7, 2025

Page 66
MassTech AI Hub Data Commons Collaborative Consulting Services RFP
RFP No: 2026-JAII-01 | October 7, 2025
Use or disclosure of data contained on this page is subject to the restriction on the title page.

Massachusetts Technology Collaborative

Grant Agreement
Standard Budget and Invoice Template
Invoiced Costs

Line Number																			
1	**Budgeting**																		
2	Applicant Information																		
3	Applicant: Archetype Consulting Inc.				Procurement No.: 2026-JAII-01														
4	Budget Period: Dec 2025 - Sep 2026				Title of Proposed Project:														
5	Address: 180 Canal St. Suite 600 Boston, MA 02114-1804				MassTech AI Hub Data Commons Collaborative Consulting Serivces														
6																			
7																			
8					MTC Funding Requested:														
9					Total Project Cost:														
10	Cost Elements				Total Project	Invoice 1		Invoice 2		Invoice 3		Invoice 4		Cumulative					
11						Period: MM/DD/YYYY-MM/DD/YYYY		Period: MM/DD/YYYY-MM/DD/YYYY		Period: MM/DD/YYYY-MM/DD/YYYY		Period: MM/DD/YYYY-MM/DD/YYYY		Period: MM/DD/YYYY-MM/DD/YYYY					
12							Amount		Amount		Amount		Amount	Expended		Remaining			
13	I. Direct Labor																		
14			*HRS*	*Rate*		*HRS*		*HRS*		*HRS*		*HRS*							
15	Mark Caradarelli	Project Manager, Sr.	1590	$250	$397,500														
16	Canon Pattillo	AI Governance Architect	830	$270	$224,100		$ -		$ -		$ -		$ -	$ -		$ 224,100.00			
17	Augusto Rosa	Lead Soultion Architect	1590	$275	$437,250		$ -		$ -		$ -		$ -	$ -		$ 437,250.00			
18	Jim Waldrop	Advisory Lead	1590	$300	$477,000		$ -		$ -		$ -		$ -	$ -		$ 477,000.00			
19	Dmitriy Ismaylov	Data Platform Engineer	1520	$225	$342,000		$ -		$ -		$ -		$ -	$ -		$ 342,000.00			
20	Glen Trudeau	DevOps/Platform Engineer	1520	$225	$342,000		$ -		$ -		$ -		$ -	$ -		$ 342,000.00			
21	Goktug Erkan	Security Architect	800	$210	$168,000														
22																			
23							$ -		$ -		$ -		$ -	$ -		$ -			
24																			
25	Total Direct Labor				$2,387,850	0	$ -	0	$ -	0	$ -	0	$ -	$ -		$ 1,822,350.00			
26																			
27	II. Direct Labor Fringe Cost	Rate (%):			$0		$ -		$ -		$ -		$ -	$ -		$ -			
28																			
29	II. Subcontractors/Consultants/Contracted Staffing																		
30	*Firm name/title or role/type of services to be provided*		*HRS*	*Rate*		*HRS*		*HRS*		*HRS*		*HRS*							
31	SMX/ UX Designer		400	$190	$76,000		$ -		$ -		$ -		$ -	$ -		$ 76,000.00			
32	SMX/WebDevelopment Engineer		800	$195	$156,000		$ -		$ -		$ -		$ -	$ -		$ 156,000.00			
33	SMX/ ServiceNow Developer		1520	$200	$304,000		$ -		$ -		$ -		$ -	$ -		$ 304,000.00			
34					$0		$ -		$ -		$ -		$ -	$ -		$ -			
35					$0	5	$ -		$ -		$ -		$ -	$ -		$ -			
36					$0		$ -		$ -		$ -		$ -	$ -		$ -			
37	Total Subcontractors/Consultants				$536,000	5	$ -	0	$ -	0	$ -	0	$ -	$ -		$ 536,000.00			
38																			
39	III. Direct Materials (provide description and costs basis)																		
40	Curriculum Development & Training				$0											$0		$0	
41					$0											$0		$0	
42					$0											$0		$0	
43					$0											$0		$0	
44					$0											$0		$0	
45					$0											$0		$0	
46					$0											$0		$0	
47	Total Direct Materials				$0		$0		$0		$0		$0	$0		$0			
48																			
49	IV. Travel: **We anticpte no travel during execution**				$0											$0		$0	
50																			
51	V. Other Direct Costs *(list type and description such as venue rental costs and cost basis)*																		
52	Snowflake (Annual Estimate)				$193,000											$ -		$ 193,000.00	
53	AWS (Annual Estimate)				$100,000											$ -		$ 100,000.00	
54	ServiceNow / Data.World (Annual Estimate)				$234,850											$ -		$ 234,850.00	
55	Venue Rental Cost (Cost incurred during the 10-months)				$6,000											$ -		$ 6,000.00	
56					$0											$ -		$ -	
57	Total Other Direct Costs				$533,850		$ -		$ -		$ -		$ -	$ -		$ 533,850.00			
58																			
59	VI. General & Administrative Expense/Overhead	Rate (%):			$0		$ -		$0		$0		$0	$ -		$0			
60																			
61	**Total PROJECT COSTS**				**$3,457,700**		**$0**		**$0**		**$0**		**$0**	**$0**		**$2,892,200**			

LEGEND:

Formula cell
Data entry cell

This section is for post-contract invoicing only. Do not complete with proposal.